Dylan D. Schmorrow
Cali M. Fidopiastis (Eds.)

LNAI 8027

Foundations of Augmented Cognition

7th International Conference, AC 2013
Held as Part of HCI International 2013
Las Vegas, NV, USA, July 2013, Proceedings

Springer

Lecture Notes in Artificial Intelligence 8027

Subseries of Lecture Notes in Computer Science

Dylan D. Schmorrow Cali M. Fidopiastis (Eds.)

Foundations of Augmented Cognition

7th International Conference, AC 2013
Held as Part of HCI International 2013
Las Vegas, NV, USA, July 21-26, 2013
Proceedings

Springer

Volume Editors

Dylan D. Schmorrow
Office of the Assistant Secretary of Defense (Research and Engineerig)
4800 Mark Center Drive, Suite 17E08, Alexandria, VA 22314, USA
E-mail: dylan.schmorrow@osd.mil

Cali M. Fidopiastis
University of Alabama at Birmingham
336 SHPB, 1530 3rd Avenue South, Birmingham, AL 35294, USA
E-mail: cfidopia@uab.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-39453-9 e-ISBN 978-3-642-39454-6
DOI 10.1007/978-3-642-39454-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013942179

CR Subject Classification (1998): H.5, K.3, H.3, K.4, H.1, I.2, I.6, J.3, J.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The 15th International Conference on Human–Computer Interaction, HCI International 2013, was held in Las Vegas, Nevada, USA, 21–26 July 2013, incorporating 12 conferences / thematic areas:

Thematic areas:

- Human–Computer Interaction
- Human Interface and the Management of Information

Affiliated conferences:

- 10th International Conference on Engineering Psychology and Cognitive Ergonomics
- 7th International Conference on Universal Access in Human–Computer Interaction
- 5th International Conference on Virtual, Augmented and Mixed Reality
- 5th International Conference on Cross-Cultural Design
- 5th International Conference on Online Communities and Social Computing
- 7th International Conference on Augmented Cognition
- 4th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- 2nd International Conference on Design, User Experience and Usability
- 1st International Conference on Distributed, Ambient and Pervasive Interactions
- 1st International Conference on Human Aspects of Information Security, Privacy and Trust

A total of 5210 individuals from academia, research institutes, industry and governmental agencies from 70 countries submitted contributions, and 1666 papers and 303 posters were included in the program. These papers address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The papers accepted for presentation thoroughly cover the entire field of Human–Computer Interaction, addressing major advances in knowledge and effective use of computers in a variety of application areas.

This volume, edited by Dylan D. Schmorrow and Cali M. Fidopiastis, contains papers focusing on the thematic area of Augmented Cognition, and addressing the following major topics:

- Augmented Cognition in Training and Education
- Team Cognition
- Brain Activity Measurement
- Understanding and Modeling Cognition
- Cognitive Load, Stress and Fatigue
- Applications of Augmented Cognition

The remaining volumes of the HCI International 2013 proceedings are:

- Volume 1, LNCS 8004, Human–Computer Interaction: Human-Centred Design Approaches, Methods, Tools and Environments (Part I), edited by Masaaki Kurosu
- Volume 2, LNCS 8005, Human–Computer Interaction: Applications and Services (Part II), edited by Masaaki Kurosu
- Volume 3, LNCS 8006, Human–Computer Interaction: Users and Contexts of Use (Part III), edited by Masaaki Kurosu
- Volume 4, LNCS 8007, Human–Computer Interaction: Interaction Modalities and Techniques (Part IV), edited by Masaaki Kurosu
- Volume 5, LNCS 8008, Human–Computer Interaction: Towards Intelligent and Implicit Interaction (Part V), edited by Masaaki Kurosu
- Volume 6, LNCS 8009, Universal Access in Human–Computer Interaction: Design Methods, Tools and Interaction Techniques for eInclusion (Part I), edited by Constantine Stephanidis and Margherita Antona
- Volume 7, LNCS 8010, Universal Access in Human–Computer Interaction: User and Context Diversity (Part II), edited by Constantine Stephanidis and Margherita Antona
- Volume 8, LNCS 8011, Universal Access in Human–Computer Interaction: Applications and Services for Quality of Life (Part III), edited by Constantine Stephanidis and Margherita Antona
- Volume 9, LNCS 8012, Design, User Experience, and Usability: Design Philosophy, Methods and Tools (Part I), edited by Aaron Marcus
- Volume 10, LNCS 8013, Design, User Experience, and Usability: Health, Learning, Playing, Cultural, and Cross-Cultural User Experience (Part II), edited by Aaron Marcus
- Volume 11, LNCS 8014, Design, User Experience, and Usability: User Experience in Novel Technological Environments (Part III), edited by Aaron Marcus
- Volume 12, LNCS 8015, Design, User Experience, and Usability: Web, Mobile and Product Design (Part IV), edited by Aaron Marcus
- Volume 13, LNCS 8016, Human Interface and the Management of Information: Information and Interaction Design (Part I), edited by Sakae Yamamoto
- Volume 14, LNCS 8017, Human Interface and the Management of Information: Information and Interaction for Health, Safety, Mobility and Complex Environments (Part II), edited by Sakae Yamamoto
- Volume 15, LNCS 8018, Human Interface and the Management of Information: Information and Interaction for Learning, Culture, Collaboration and Business (Part III), edited by Sakae Yamamoto
- Volume 16, LNAI 8019, Engineering Psychology and Cognitive Ergonomics: Understanding Human Cognition (Part I), edited by Don Harris
- Volume 17, LNAI 8020, Engineering Psychology and Cognitive Ergonomics: Applications and Services (Part II), edited by Don Harris
- Volume 18, LNCS 8021, Virtual, Augmented and Mixed Reality: Designing and Developing Augmented and Virtual Environments (Part I), edited by Randall Shumaker

- Volume 19, LNCS 8022, Virtual, Augmented and Mixed Reality: Systems and Applications (Part II), edited by Randall Shumaker
- Volume 20, LNCS 8023, Cross-Cultural Design: Methods, Practice and Case Studies (Part I), edited by P.L. Patrick Rau
- Volume 21, LNCS 8024, Cross-Cultural Design: Cultural Differences in Everyday Life (Part II), edited by P.L. Patrick Rau
- Volume 22, LNCS 8025, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Healthcare and Safety of the Environment and Transport (Part I), edited by Vincent G. Duffy
- Volume 23, LNCS 8026, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Human Body Modeling and Ergonomics (Part II), edited by Vincent G. Duffy
- Volume 25, LNCS 8028, Distributed, Ambient and Pervasive Interactions, edited by Norbert Streitz and Constantine Stephanidis
- Volume 26, LNCS 8029, Online Communities and Social Computing, edited by A. Ant Ozok and Panayiotis Zaphiris
- Volume 27, LNCS 8030, Human Aspects of Information Security, Privacy and Trust, edited by Louis Marinos and Ioannis Askoxylakis
- Volume 28, CCIS 373, HCI International 2013 Posters Proceedings (Part I), edited by Constantine Stephanidis
- Volume 29, CCIS 374, HCI International 2013 Posters Proceedings (Part II), edited by Constantine Stephanidis

I would like to thank the Program Chairs and the members of the Program Boards of all affiliated conferences and thematic areas, listed below, for their contribution to the highest scientific quality and the overall success of the HCI International 2013 conference.

This conference could not have been possible without the continuous support and advice of the Founding Chair and Conference Scientific Advisor, Prof. Gavriel Salvendy, as well as the dedicated work and outstanding efforts of the Communications Chair and Editor of HCI International News, Abbas Moallem.

I would also like to thank for their contribution towards the smooth organization of the HCI International 2013 Conference the members of the Human–Computer Interaction Laboratory of ICS-FORTH, and in particular George Paparoulis, Maria Pitsoulaki, Stavroula Ntoa, Maria Bouhli and George Kapnas.

May 2013

Constantine Stephanidis
General Chair, HCI International 2013

Organization

Human–Computer Interaction

Program Chair: Masaaki Kurosu, Japan

Jose Abdelnour-Nocera, UK
Sebastiano Bagnara, Italy
Simone Barbosa, Brazil
Tomas Berns, Sweden
Nigel Bevan, UK
Simone Borsci, UK
Apala Lahiri Chavan, India
Sherry Chen, Taiwan
Kevin Clark, USA
Torkil Clemmensen, Denmark
Xiaowen Fang, USA
Shin'ichi Fukuzumi, Japan
Vicki Hanson, UK
Ayako Hashizume, Japan
Anzai Hiroyuki, Italy
Sheue-Ling Hwang, Taiwan
Wonil Hwang, South Korea
Minna Isomursu, Finland
Yong Gu Ji, South Korea
Esther Jun, USA
Mitsuhiko Karashima, Japan
Kyungdoh Kim, South Korea
Heidi Krömker, Germany
Chen Ling, USA
Yan Liu, USA
Zhengjie Liu, P.R. China
Loïc Martínez Normand, Spain
Chang S. Nam, USA
Naoko Okuizumi, Japan
Noriko Osaka, Japan
Philippe Palanque, France
Hans Persson, Sweden
Ling Rothrock, USA
Naoki Sakakibara, Japan
Dominique Scapin, France
Guangfeng Song, USA
Sanjay Tripathi, India
Chui Yin Wong, Malaysia
Toshiki Yamaoka, Japan
Kazuhiko Yamazaki, Japan
Ryoji Yoshitake, Japan
Silvia Zimmermann, Switzerland

Human Interface and the Management of Information

Program Chair: Sakae Yamamoto, Japan

Hans-Jorg Bullinger, Germany
Alan Chan, Hong Kong
Gilsoo Cho, South Korea
Jon R. Gunderson, USA
Shin'ichi Fukuzumi, Japan
Michitaka Hirose, Japan
Jhilmil Jain, USA
Yasufumi Kume, Japan
Mark Lehto, USA
Hiroyuki Miki, Japan
Hirohiko Mori, Japan
Fiona Fui-Hoon Nah, USA
Shogo Nishida, Japan
Robert Proctor, USA
Youngho Rhee, South Korea
Katsunori Shimohara, Japan

Michale Smith, USA
Tsutomu Tabe, Japan
Hiroshi Tsuji, Japan
Kim-Phuong Vu, USA
Tomio Watanabe, Japan
Hidekazu Yoshikawa, Japan

Engineering Psychology and Cognitive Ergonomics

Program Chair: Don Harris, UK

Guy Andre Boy, USA
Joakim Dahlman, Sweden
Trevor Dobbins, UK
Mike Feary, USA
Shan Fu, P.R. China
Michaela Heese, Austria
Hung-Sying Jing, Taiwan
Wen-Chin Li, Taiwan
Mark A. Neerincx, The Netherlands
Jan M. Noyes, UK
Taezoon Park, Singapore
Paul Salmon, Australia
Axel Schulte, Germany
Siraj Shaikh, UK
Sarah C. Sharples, UK
Anthony Smoker, UK
Neville A. Stanton, UK
Alex Stedmon, UK
Xianghong Sun, P.R. China
Andrew Thatcher, South Africa
Matthew J.W. Thomas, Australia
Rolf Zon, The Netherlands

Universal Access in Human–Computer Interaction

Program Chairs: Constantine Stephanidis, Greece, and Margherita Antona, Greece

Julio Abascal, Spain
Ray Adams, UK
Gisela Susanne Bahr, USA
Margit Betke, USA
Christian Bühler, Germany
Stefan Carmien, Spain
Jerzy Charytonowicz, Poland
Carlos Duarte, Portugal
Pier Luigi Emiliani, Italy
Qin Gao, P.R. China
Andrina Granić, Croatia
Andreas Holzinger, Austria
Josette Jones, USA
Simeon Keates, UK
Georgios Kouroupetroglou, Greece
Patrick Langdon, UK
Seongil Lee, Korea
Ana Isabel B.B. Paraguay, Brazil
Helen Petrie, UK
Michael Pieper, Germany
Enrico Pontelli, USA
Jaime Sanchez, Chile
Anthony Savidis, Greece
Christian Stary, Austria
Hirotada Ueda, Japan
Gerhard Weber, Germany
Harald Weber, Germany

Virtual, Augmented and Mixed Reality

Program Chair: Randall Shumaker, USA

Waymon Armstrong, USA
Juan Cendan, USA
Rudy Darken, USA
Cali M. Fidopiastis, USA
Charles Hughes, USA
David Kaber, USA
Hirokazu Kato, Japan
Denis Laurendeau, Canada
Fotis Liarokapis, UK
Mark Livingston, USA
Michael Macedonia, USA
Gordon Mair, UK
Jose San Martin, Spain
Jacquelyn Morie, USA
Albert "Skip" Rizzo, USA
Kay Stanney, USA
Christopher Stapleton, USA
Gregory Welch, USA

Cross-Cultural Design

Program Chair: P.L. Patrick Rau, P.R. China

Pilsung Choe, P.R. China
Henry Been-Lirn Duh, Singapore
Vanessa Evers, The Netherlands
Paul Fu, USA
Zhiyong Fu, P.R. China
Fu Guo, P.R. China
Sung H. Han, Korea
Toshikazu Kato, Japan
Dyi-Yih Michael Lin, Taiwan
Rungtai Lin, Taiwan
Sheau-Farn Max Liang, Taiwan
Liang Ma, P.R. China
Alexander Mädche, Germany
Katsuhiko Ogawa, Japan
Tom Plocher, USA
Kerstin Röse, Germany
Supriya Singh, Australia
Hsiu-Ping Yueh, Taiwan
Liang (Leon) Zeng, USA
Chen Zhao, USA

Online Communities and Social Computing

Program Chairs: A. Ant Ozok, USA, and Panayiotis Zaphiris, Cyprus

Areej Al-Wabil, Saudi Arabia
Leonelo Almeida, Brazil
Bjørn Andersen, Norway
Chee Siang Ang, UK
Aneesha Bakharia, Australia
Ania Bobrowicz, UK
Paul Cairns, UK
Farzin Deravi, UK
Andri Ioannou, Cyprus
Slava Kisilevich, Germany
Niki Lambropoulos, Greece
Effie Law, Switzerland
Soo Ling Lim, UK
Fernando Loizides, Cyprus
Gabriele Meiselwitz, USA
Anthony Norcio, USA
Elaine Raybourn, USA
Panote Siriaraya, UK
David Stuart, UK
June Wei, USA

Augmented Cognition

Program Chairs: Dylan D. Schmorrow, USA, and Cali M. Fidopiastis, USA

Robert Arrabito, Canada
Richard Backs, USA
Chris Berka, USA
Joseph Cohn, USA
Martha E. Crosby, USA
Julie Drexler, USA
Ivy Estabrooke, USA
Chris Forsythe, USA
Wai Tat Fu, USA
Rodolphe Gentili, USA
Marc Grootjen, The Netherlands
Jefferson Grubb, USA
Ming Hou, Canada
Santosh Mathan, USA
Rob Matthews, Australia
Dennis McBride, USA
Jeff Morrison, USA
Mark A. Neerincx, The Netherlands
Denise Nicholson, USA
Banu Onaral, USA
Lee Sciarini, USA
Kay Stanney, USA
Roy Stripling, USA
Rob Taylor, UK
Karl van Orden, USA

Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management

Program Chair: Vincent G. Duffy, USA and Russia

Karim Abdel-Malek, USA
Giuseppe Andreoni, Italy
Daniel Carruth, USA
Eliza Yingzi Du, USA
Enda Fallon, Ireland
Afzal Godil, USA
Ravindra Goonetilleke, Hong Kong
Bo Hoege, Germany
Waldemar Karwowski, USA
Zhizhong Li, P.R. China
Kang Li, USA
Tim Marler, USA
Michelle Robertson, USA
Matthias Rötting, Germany
Peter Vink, The Netherlands
Mao-Jiun Wang, Taiwan
Xuguang Wang, France
Jingzhou (James) Yang, USA
Xiugan Yuan, P.R. China
Gülcin Yücel Hoge, Germany

Design, User Experience, and Usability

Program Chair: Aaron Marcus, USA

Sisira Adikari, Australia
Ronald Baecker, Canada
Arne Berger, Germany
Jamie Blustein, Canada
Ana Boa-Ventura, USA
Jan Brejcha, Czech Republic
Lorenzo Cantoni, Switzerland
Maximilian Eibl, Germany

Anthony Faiola, USA
Emilie Gould, USA
Zelda Harrison, USA
Rüdiger Heimgärtner, Germany
Brigitte Herrmann, Germany
Steffen Hess, Germany
Kaleem Khan, Canada
Jennifer McGinn, USA
Francisco Rebelo, Portugal
Michael Renner, Switzerland
Kerem Rızvanoğlu, Turkey
Marcelo Soares, Brazil
Christian Sturm, Germany
Michele Visciola, Italy

Distributed, Ambient and Pervasive Interactions

Program Chairs: Norbert Streitz, Germany, and Constantine Stephanidis, Greece

Emile Aarts, The Netherlands
Adnan Abu-Dayya, Qatar
Juan Carlos Augusto, UK
Boris de Ruyter, The Netherlands
Anind Dey, USA
Dimitris Grammenos, Greece
Nuno M. Guimaraes, Portugal
Shin'ichi Konomi, Japan
Carsten Magerkurth, Switzerland
Christian Müller-Tomfelde, Australia
Fabio Paternó, Italy
Gilles Privat, France
Harald Reiterer, Germany
Carsten Röcker, Germany
Reiner Wichert, Germany
Woontack Woo, South Korea
Xenophon Zabulis, Greece

Human Aspects of Information Security, Privacy and Trust

Program Chairs: Louis Marinos, ENISA EU, and Ioannis Askoxylakis, Greece

Claudio Agostino Ardagna, Italy
Zinaida Benenson, Germany
Daniele Catteddu, Italy
Raoul Chiesa, Italy
Bryan Cline, USA
Sadie Creese, UK
Jorge Cuellar, Germany
Marc Dacier, USA
Dieter Gollmann, Germany
Kirstie Hawkey, Canada
Jaap-Henk Hoepman, The Netherlands
Cagatay Karabat, Turkey
Angelos Keromytis, USA
Ayako Komatsu, Japan
Ronald Leenes, The Netherlands
Javier Lopez, Spain
Steve Marsh, Canada
Gregorio Martinez, Spain
Emilio Mordini, Italy
Yuko Murayama, Japan
Masakatsu Nishigaki, Japan
Aljosa Pasic, Spain
Milan Petković, The Netherlands
Joachim Posegga, Germany
Jean-Jacques Quisquater, Belgium
Damien Sauveron, France
George Spanoudakis, UK
Kerry-Lynn Thomson, South Africa

Julien Touzeau, France
Theo Tryfonas, UK
João Vilela, Portugal
Claire Vishik, UK
Melanie Volkamer, Germany

External Reviewers

Maysoon Abulkhair, Saudi Arabia
Ilia Adami, Greece
Vishal Barot, UK
Stephan Böhm, Germany
Vassilis Charissis, UK
Francisco Cipolla-Ficarra, Spain
Maria De Marsico, Italy
Marc Fabri, UK
David Fonseca, Spain
Linda Harley, USA
Yasushi Ikei, Japan
Wei Ji, USA
Nouf Khashman, Canada
John Killilea, USA
Iosif Klironomos, Greece
Ute Klotz, Switzerland
Maria Korozi, Greece
Kentaro Kotani, Japan
Vassilis Kouroumalis, Greece
Stephanie Lackey, USA
Janelle LaMarche, USA
Asterios Leonidis, Greece
Nickolas Macchiarella, USA
George Margetis, Greece
Matthew Marraffino, USA
Joseph Mercado, USA
Claudia Mont'Alvão, Brazil
Yoichi Motomura, Japan
Karsten Nebe, Germany
Stavroula Ntoa, Greece
Martin Osen, Austria
Stephen Prior, UK
Farid Shirazi, Canada
Jan Stelovsky, USA
Sarah Swierenga, USA

HCI International 2014

The 16th International Conference on Human–Computer Interaction, HCI International 2014, will be held jointly with the affiliated conferences in the summer of 2014. It will cover a broad spectrum of themes related to Human–Computer Interaction, including theoretical issues, methods, tools, processes and case studies in HCI design, as well as novel interaction techniques, interfaces and applications. The proceedings will be published by Springer. More information about the topics, as well as the venue and dates of the conference, will be announced through the HCI International Conference series website: `http://www.hci-international.org/`

General Chair
Professor Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
Email: `cs@ics.forth.gr`

Table of Contents

Augmented Cognition in Training and Education

Team Cognition

Brain Activity Measurement

Understanding and Modelling Cognition

Cognitive Load, Stress and Fatigue

Applications of Augmented Cognition

Part I

Augmented Cognition in Training and Education

Intuitive Sensemaking: From Theory to Simulation Based Training

Kathleen Bartlett, Margaret Nolan, and Andrea Marraffino

MESH Solutions, LLC – A DSCI Company, Orlando, FL, USA
{kbartlett,mnolan,amarraffino}@mesh.dsci.com

Abstract. The concept of *sensemaking* has become a prominent component of military operations in ambiguous environments. Sensemaking, in general, describes the process of pattern recognition, semantic formulation, anticipation, and holistic understanding and supports sociocultural situation assessment, anomaly detection, and anticipatory thinking. This skill enables intuitive experts to rapidly draw accurate conclusions based on cues that others cannot discern or to attend to the most important cues, based on experience. Simulation-based training can enhance and accelerate the ability to recognize and analyze cues and patterns by translating the unconscious, automatic monitoring and integration practiced by experts into a conscious cognitive process that we call intuitive sensemaking. We describe an Office of Naval Research project, currently in development, intended to effectively train previously ambiguous advanced cognitive skills such as intuition-informed sensemaking. With training, teams of military personnel should see increases in cohesiveness, sociocultural situation assessment, anomaly detection, and anticipatory thinking.

Keywords: Sensemaking, Intuition, Simulation-Based Training, Human-Computer Interaction (HCI), Expertise, Implicit Learning.

1 Introduction: Sensemaking and Intuition

Sensemaking describes the ability to explain data that are sparse, noisy, and uncertain (Moore, 2011) when assessing a situation. Sociologist Karl Weick, one of the first academics to define sensemaking as it relates directly to complex operational environments, contended that the ability to construct a coherent and shared explanation for events and circumstances enables operational functioning during periods of great uncertainty (Weick, 1993). In the past, this type of intelligence gathering generally consisted of locating a known entity or a specific target; today's battlefields, however, require an additional skill set: looking for (and judging the significance of) undefined activities or transactions. Frequently, observers must scan complex, ambiguous settings and groups of diverse, unpredictable people to assess threats and determine necessary actions. They must make sense of situations that include large numbers of relatively small actors responding to a shifting set of situational factors (Moore, 2011). Predicting and anticipating the actions of these players and the directional

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 3–10, 2013.

shifts of surrounding circumstances requires enhanced observational and sensemaking training that accelerates the acquisition of expertise and fosters the development of intuition.

Individuals engage in sensemaking under conditions of equivocality and uncertainty (Weick, 1979, 1993), and their expectations and motivations affect this process, since individuals vary in how they construct ethical issues and make intuitive judgments about those constructions (Sonenshein, 2007). While some people seem to be more naturally intuitive than others, recent work suggests that the process of intuition rests on an unconscious awareness, valuation, and integration of cues that shape decisions and judgments, and in experts, those perceptual observations may reach a level of automaticity (e.g., Dervin, B., 1983; Klein, Moon, & Hoffman, 2006b; Thurlow & Mills, 2009; Betsch, 2008; Dane & Pratt, 2007). When people "know" without knowing how they know, their conscious awareness may have been influenced by an unconscious monitoring of patterns and anomalies (e.g., Claxton, 2000; Simons & Chabris, 2010). For example, profound decisions and actions that save firefighters in potentially catastrophic situations can most likely be credited to implicit processing of important environmental cues. Decades of research on implicit learning have shown that our brains possess an array of mechanisms for automatically extracting information from the environment without our awareness (Reber, 2008). This skill enables intuitive experts to rapidly draw accurate conclusions based on cues that others cannot discern or to attend to the most important cues, given conditions and context of a situation, based on experience.

The human brain has two distinct information processing systems: one conscious and deliberative and the other unconscious and intuitive. Intuition is rooted in the unconscious information processing system, as are related inputs of implicit attitudes and goals (Hassin, Uleman, & Bargh, 2005). This intuitive processing creates the moment of intuition, the experience of knowing, without knowing how that knowledge came to be. Dane and Pratt (2007) offer this definition: intuitions are "affectively charged judgments that arise through rapid, non-conscious, and holistic associations" (Dane & Pratt, 2007). Betsch (2008) provides a definition of the three core components of intuition: "Intuition is a process of thinking. The input to this process is mostly provided by knowledge stored in long-term memory that has been primarily acquired via associative learning. The input is processed automatically and without conscious awareness. The output of the process is a feeling that can serve as a basis for judgments and decisions." Thus, while experts may attribute their advanced awareness (e.g., sense of danger before an explosion) to intuition, at an unconscious level, they most likely had mentally observed, analyzed, and decided how to act without recognizing that cognitive process. Intuition typically emerges with no awareness of the mental events leading to it, which fits with our conjecture that implicit memory is critical in producing trustworthy intuition. The results of this implicit learning often appear as an intuition or a "sixth sense" about the current situation. Development of expertise that achieves this level of automaticity takes time, but simulation-based training can accelerate that process.

2 Can Sensemaking and Intuition Be Trained?

Research suggests that fast, affect-rich intuitions frequently drive individuals' behavior (Loewenstein, 1996). Subject-matter experts routinely use their intuitive abilities to help them make decisions and judgments (Hodgkinson et al., 2009), and intuition is "critical to effective decision making in many settings" (Salas et al., 2009, p. 2). Emerging theory further proposes the existence of expertise-based intuition, a form of intuition rooted in domain-specific expertise that experts can learn to constructively employ in support of their decision making, sensemaking, and other cognitive processes (Salas et al., 2009). This suggests that intuition, like other cognitive mechanisms, can improve through experience, deliberate practice, and a variety of specialized training interventions. It also means that it may be possible to decrease the time required to effectively use intuition to drive decisions in situations where one has not had the requisite time required to become a domain expert. In other words, we may be able to artificially enhance intuitive decision making skills at a more rapid pace than previous research has suggested (Eriksson, 1996) by using specific training techniques within implicit learning environments.

Given the influence of intuition on cognitive performance, and considering the recent evidence that individuals can intentionally improve their intuitive skills, the military community's interest in intuition has grown. Because intuition and psychosocial skills complement one another, training for intuitive processing might be conveyed in the learning context of sociocultural pattern recognition, anticipation, and interaction. In that context, intuitive processing can enhance the discernment and interpretation of subtle sociocultural cues and patterns, and supports the need to enhance military personnel's sociocultural abilities. Moreover, such blended instruction on these topics may help engender the generalizable sociocultural competencies that Marines and Sailors need to excel in any operational environment.

Despite its subconscious facets, training and education can enhance individuals' intuitive capacities (Salas et al., 2009). Classically, experts build their intuitive skills through experience and implicit learning (Agor, 1989; Harper, 1989; Klein, 1998). They learn to regulate their intuitive feelings by actively seeking feedback (Hogarth, 2001), and they selectively attend to intuitive thoughts based upon the characteristics of the problem space (Salas et al., 2009). Fortunately, accelerated acquisition of domain experience and the development of intuition-related skills can be facilitated through deliberate practice, critical self-appraisal, and candid feedback (Hodgkinson, 2009). A validated training program for intuition could help military personnel improve their access to, and appropriate use of, intuition. For example, through implicit learning, situated training, deliberate practice, self-critique, and metacognitive instruction, warfighters could enhance their intuition-informed pattern recognition capabilities, learn to more rapidly and efficaciously conduct intuition-informed situation assessments, and gain regulatory skills to more deliberately control their intuitive processing.

Specifically, simulation-based training can enhance and accelerate the ability to recognize and analyze cues and patterns by translating the unconscious, automatic

monitoring and integration practiced by experts into a conscious cognitive process that we call intuitive sensemaking. However, the challenge consists of taking empirically-driven findings about intuition, based on models and theories, best practices in simulation-based training, and what is known about instructional strategies and feedback, to design effective scenarios that will stimulate and train the development of these unconscious intuitive functions.

For simulation-based training, the scenario design provides the context for the training; it defines the capabilities of the simulation system that should be utilized to create the required conditions and cues, and suggests the instructional strategies that provide the best method to deliver the training and the performance feedback. Will it be simple or complex? Is a part-task trainer that adds layers as expertise improves the best option to train the objectives? Is immediate feedback or an after-action review the best way to emphasize learning points? Should feedback be embedded in the scenario design or mediated by a live instructor? Design decisions will also involve "visual noise" and temporal markers, as well as spatial issues of proximity, juxtaposition, and foreshadowing. How can we design a richly cued scenario that will enable or direct a search, and what are the salient cues important to the overall training objective that will need to be detected?

Event-based training featuring novel, unexpected situations offers friction points to stimulate decision making under time or mission constraints and to provide triggers for other courses of action; timely feedback can direct attention to cues and patterns that are missed. Simulations offer a chance to rehearse actions and thinking, but capturing the thinking and knowing that occurs below conscious awareness presents another challenge. Simulations also provide a means for repetition of the same or similar scenarios, increasing the complexity in a chained strategy, and provide the means for transfer of training to a novel scenario. Performance measurement is generally comprised of observable outcomes and courses of action taken based on decision(s) made in support of an objective. Simulation-based technologies can provide capture of voice and video recordings, resources and assets used, and digital information exchanged to help build metrics. Measurement of physiological functions (e.g., via EEG, eye tracking) provide data that may help identify what is being observed and when, for how long, in what sequence, and how that information influences decision making and outcomes. Correlating those data with trainee verbal protocol and demonstrable results from actions taken can provide insights into implicit decision making and how best to employ training feedback to uncover the trainees' strategies, such as how or if they took into account cue characteristics, multiple criteria, sequence of visual acquisition of cues, etc.

Scenario-based simulation training to accelerate the process of automaticity will need to develop the skills of an expert via intense and intentional practice with specific feedback that impacts the learner in an emotional, highly connotative way. Feedback delivery options must also be explored, since allowing learners to make errors can increase problem solving abilities and enhance the emotional impact of the experience, whereas providing delayed feedback, or after-action review, may create a disconnect between action and consequence (or may lead to negative training). Consensus among

experts on what constitutes "good" decisions, and therefore actions, will need to be addressed to ensure consistent training objectives and sound scenario and metric design. Decomposition of the decisions and "micro" decisions that could or should be made within the context of a given scenario is paramount for simulation design.

2.1 PercepTS: Immersive Technologies to Enhance Intuitive Decision Making

We theorize that, after exploring factors and options related to simulation-based training, military researchers will have the ability to effectively train previously ambiguous advanced cognitive skills such as intuition-informed sensemaking. With training, teams of military personnel should see increases in cohesiveness, sociocultural situation assessment, anomaly detection, and anticipatory thinking.

Toward this end, the Office of Naval Research (ONR) stood up the Perceptual Training Systems and Tools (PercepTS) program to explore immersive methodologies and technologies for improving the training of sensemaking/perceptual knowledge, skills, and abilities (KSAs) in operational environments. This work includes development of an approach to enhance the decision-making skills of military personnel by investigating a range of cognitive training approaches, situated in the context of urban sociocultural sensemaking.

On-going PercepTS work seeks to develop an actionable framework of perceptual competencies and training strategies for military use. Based on this framework and the instructional strategies identified therein, a perceptual skils Program of Instruction (POI) suitable for implementation by military instructors could be developed under future projects, tested, and packaged as advanced instructional strategies for future use in adaptive implicit training systems. Given the current (and likely future) emphasis on Stabilization, Security, Transition, and Reconstruction Operations (SSTRO) and other socioculturally situated operations, intuition research is being conducted and developed for psychosocial skills training. In particular, the Virtual Observation Platform, a simulation-based trainer in development under PercepTS, could act as a test site for implicit learning techniques to enhance intuitive decision making. This simulation-based training approach offers opportunities for practice in sensemaking activities, via anomaly detection among patterns of human behavior. Repeated opportunities for practice can accelerate the development of expertise, wherein the sensemaking process of informed observation, analysis, and action becomes intuitive. Development of intuitive sensemaking will provide learners with the ability to solve ill-defined problems and make sound, complex decisions in uncertain, socially complicated operating environments.

3 Additional On-going and Future Work

Future work in this area might explore the use of adaptive training interventions, including neurophysiologically informed adaptive instructional systems currently being investigated by other researchers. For example, under DARPA's Warfighter Intuition

effort, using a high-density EEG and post hoc analyses, researchers identified medial orbital frontal electrical responses that seem to correlate with presence of intuition (Luu et al., 2010). The cumbersome technology and time-consuming analyses prohibit this technology from being deployed in the near term; however, it is reasonable to believe that a neurophysiologically informed adaptive training system could support intuition instruction.

In Spring of 2013, work will begin on an ONR-funded program to address four areas of a Basic Research Challenge: "Enhancing Intuitive Decision Making Through Implicit Learning." A world-class team of researchers will combine the talents of the groundbreaking cognitive and neuroscience university laboratories at Northwestern University (NWU), Massachusetts Institute of Technology (MIT), and University of California, Los Angeles (UCLA) with teams of human systems modeling and simulation scientists from Charles River Associates, and Defense Group Inc., all led by MESH Solutions/DSCI to contribute to the Intuitive Sensemaking Interactive Simulation (ISIS) program. This unique team will collaborate to research, develop models, make recommendations, and test advanced instructional strategies for use in adaptive implicit scenario/simulation based training system.

The phased approach of the program will include parallel Neuroimaging experiments conducted at the IMHRO Staglin Center for Cognitive Neuroscience at UCLA and the Center for Translational Imaging at NWU to investigate multiple brain processes in order to gain greater understanding of intuitive decision making. Computational models will be developed and used for both data analysis and to inform recommendations for instructional strategies. Results in the form of modeling predictions and training recommendations will be used to drive simulation-based training experimentation and to explore the training effectiveness, and validation of, the resulting guidelines and strategies.

Current and future work will need to articulate and refine possible training strategies that could be triggered or enhanced by neurophysiological inputs and to recommend adaptive training strategies that a developer could implement into an intuition intelligent tutor, once the corresponding sensing technologies reach sufficient maturity. The overarching goal of this work is to advance capabilities for enhancing the intuitive decision-making skills of military personnel by investigating a range of intuition training approaches, situated in the context of urban sociocultural sensemaking.

Acknowledgement. This work was supported, in part, by the Office of Naval Research project N00014-11-C-0193, Perceptual Training Systems and Tools (PercepTS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

1. Agor, W.H. (ed.): Intuition in Organizations: Leading and Managing Productively. Sage Publications, Newbury Park (1989)
2. Betsch, T.: Intuition in Judgment and Decision Making. Taylor and Francis Publishers, New York (2008)
3. Claxton, G.: The anatomy of intuition. In: Atkinson, T., Claxton, G. (eds.) TheIntuitive Practitioner. Open University Press, Buckingham (2000)
4. Dane, E., Pratt, M.: Exploring intuition and its role in managerial decision making. Academy of Management Review 32(1), 33–64 (2007)
5. Dervin, B.: An overview of sense-making research: Concepts, methods and results. Paper Presented at the Annual Meeting of the International Communication Association, Dallas, TX (1983)
6. Ericsson, K.A., Lehmann, A.C.: Expert and exceptional performance: Evidence of maximal adaptation to task constraints. Annual Review of Psychology 47, 273–305 (1996)
7. Harper, S.C.: Intuition: What separates executives from managers. In: Agor, W.H. (ed.) Intuition in Organizations, pp. 111–124. Sage Publications, Newbury Park (1989)
8. Hassin, R.R., Uleman, J.S., Bargh, J.A. (eds.): The New Unconscious. Oxford University Press, Oxford (2005)
9. Hodgkinson, G.P.: Intuition in Organizational Decision making. Keynote address to the IST Workshop, February 4-5. University of Central Florida (2009)
10. Hogarth, R.M.: Educating Intuition. University of Chicago Press, Chicago (2001)
11. Klein, G.: Sources of Power: How People Make Decisions. MIT Press, Cambridge (1998)
12. Klein, G., Moon, B., Hoffman, R.: Making sense of sensemaking 1: Alternative perspectives. Intelligent Systems 21(4) (2006a)
13. Klein, G., Moon, B., Hoffman, R.F.: Making sense of sensemaking Ii: a macrocognitive model. IEEE Intelligent Systems 21(5), 88–92 (2006b)
14. Lowenstein, R.: When Genius Failed: The Rise and Fall of Long-Term Capital Management. Random House Trade, New York (2000)
15. Luu, P., Geyer, A., Fidopiastis, C., Campbell, G., Wheeler, T., et al.: Reentrant Processing in Intuitive Perception. PLoS One 5(3), e9523 (2010)
16. Moore, D.T.: Sensemaking: A Structure for an Intelligence Revolution. Clift Series on the Intelligence Profession. National Defense Intelligence College (2011)
17. Reber, P.J.: Cognitive neuroscience of declarative and non-declarative memory (2008)
18. Guadagnoli, M., de Belle, S., Etnyre, B., Polk, T., Benjamin, A. (eds.): Parallels in Learning and Memory, 113–123
19. Salas, E., Rosen, M.A., Diaz Granados, D.: Expertise-Based Intuition and Decision Making in Organizations. Journal of Management, 1–31 (2009)
20. Schatz, S., Wray, R., Folsom-Kovarik, J.T., Nicholson, D.: Adaptive Perceptual Training in a Virtual Environment. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, Florida, December 3-6 (2012a)
21. Schatz, S., Nicholson, D.: Perceptual Training for Cross-Cultural Decision Making. In: Nicholson, D.M. (ed.) Advances for Design in Cross-Cultural Activities Part I, pp. 3–12. CRC Press, Boca Raton (2012)
22. Schatz, S., Folsom-Kovarik, J.T., Barlett, K., Wray, R., Solina, D.: Archetypal Patterns of Life for Military Training Simulations. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, Florida, December 3-6 (2012)

23. Simons, D., Chabris, C.: The Trouble with Intuition. The Chronicle of Higher Education (2010)
24. Sonenshein, S.: The role of construction, intuition, and justification in responding to ethical issues at work: The sensemaking-intuition model. Academy of Management Review 32(4), 1022–1040 (2007)
25. Thurlow, A., Mills, J.: Change, talk and sensemaking. Journal of Organizational Change Management 22(5), 459–579 (2009)
26. Weick, K.E.: The Social Psychology of Organizing, 2nd edn. McGraw-Hill, New York (1979)
27. Weick, K.: The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. Administrative Science Quarterly (38), 628–652 (1993)

Using Simulation Based Training Methods for Improved Warfighter Decision Making

Perakath Benjamin, Paul Koola, Kumar Akella, Michael Graul, and Michael Painter

Knowledge Based Systems, Inc., USA
pbenjamin@kbsi.com

Abstract. Few efforts have greater significance to our warfighting capability than those aimed at dramatically improving the skills, knowledge, and experience of military decision makers. The research and technology ideas presented in this paper are motivated by need to improve the quality of decision makers through the design of innovative training technology for military decision makers. This paper describes an adaptive simulation based training approach to improve the effectiveness of warfighter decision making. The paper describes (i) a method for adaptive simulation based training; (ii) a mission-driven approach to measure trainee performance based on carefully designed metrics; and (iii) an automation support architecture for adaptive simulation based training. Examples are provided throughout the paper to illustrate key research ideas.

Keywords: Simulation Based Training, Warfighter Decision Making, Adaptive Training, Mission Driven Performance Measurement.

1 Motivations

This section summarizes important technology gaps in the area of simulation-based training to enhance warfighter decision making effectiveness. The central problem is the inability to rapidly refine and adapt simulation based training content to address focused training needs. Currently, simulation based training content is painstakingly handcrafted by subject matter experts (SMEs) and this content is not maintained in a manner that facilitates rapid change. Moreover, simulation based training systems do not provide mechanisms for automatically determining training content changes based on the analysis of measured student performance. Current simulation-based training systems lack the ability to efficiently adapt the current state of a scenario to a desired state that will address the training goals. Consider the simulation-based Military Operations on Urban Training (MOUT) infantry training exercise in which the goal is to detect and eliminate a sniper. If, during the training, the trainee constructs a smoke-screen or exits his vantage point, these actions serve to render the scenario ineffective for the intended goal. To use another example, in an air combat exercise focused on increasing threat awareness in the presence of enemy radar sites, the trainee's departure from the radar site area renders the scenario ineffective for the

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 11–20, 2013.

exercise's "increase threat awareness" goals. Under current simulation-based training systems, the instructor would have to issue a request to the simulator operator in order to create a set of new Computer Generated Forces (CGFs) for the trainee in order to meet the goals of the exercise. Such a request has an unacceptable response time, especially towards the end of the exercise. It is often the case that the instructor defers the unfulfilled training goal to a future simulation exercise. This carries the risk that the same course of events would ensue even in future exercises. Adaptive simulation content generation methods may be used to automate the generation of new training drills and scenarios within seconds of sub-optimal trainee actions. In the first example, a new CGF action (a simulation 'drill') will be automatically inserted to bring another bandit to replace the originally defeated bandit. In the second example, a new radar site would be automatically inserted into the area in which the trainee strayed.

1.1 Lack of Knowledge Capture Methods and Tools

There exists a technology gap surrounding effective methods for capturing and maintaining critical training event data such as training goals, trainee decisions, trainee performance, etc. Scenario-based training provides an advanced framework for decision makers to be exposed to real tasks in a systematic way. It is also a practical approach because it facilitates the move toward an adaptive training paradigm, in which new incidents may be defined and deployed during the training exercise. Scenario-based training is composed of six main steps executed in a closed cycle: (i) Skill Inventory/Performance Data, (ii) Learning Objectives/Competencies, (iii) Scenario Events/Scripts, (iv) Performance Measures/Standards, (v) Performance Diagnosis, and (vi) Feedback and Debrief [1]. Under scenario-based training, trainers are responsible for monitoring trainees, providing feedback, diagnosing deficiencies and performing remediation. However, correct execution of all of these tasks represents a huge task overload on trainers that are involved in the scenario-based training process, and existing scenario definition tools lack comprehensive knowledge capture and knowledge management capabilities to compensate for this overload. Further, relations between the objects of the scenario and information about the relations that exist between objects in the scenario are not maintained, thereby making it nearly impossible to perform automated after-action review and historical analysis.

1.2 Lack of Knowledge Capture and Reuse Technology for Training

There exists a void in the availability of methods and tools for capturing and reusing training information from recurring training events. For example, tools are necessary to maintain information about the students participating in the training, their role types, their association to other training events, what roles they played in those events, their performance in those events, characteristics of their training regimen that they felt were most influential in their performance, etc. Further, mechanisms are needed to capture and use training lessons learned over recurring exercises.

1.3 Lack of Learning Mechanisms for Training Systems

Absent are adaptive automation mechanisms to improve the quality and content of training over time. Learning Management System (LMS) technologies must allow the students to provide scenario enhancements based upon their experience. In essence, the scenario definition system must "learn" about or adapt to new aspects of the training environment or objects in the system and allow the scenario developer to utilize these new facets in the generation of new scenarios.

1.4 Paper Outline

This paper describes simulation based methods and automation mechanisms that seek to address the above challenges. First, we will describe an ontology for adaptive simulation based training that provides a conceptual foundation for the adaptive training method. Next, we outline an adaptive simulation based training method. A mission-driven approach to measure performance is described. A summary of an automation architecture for adaptive simulation based training is then presented. Finally, the paper summarizes the benefits of our adaptive simulation based training method and outlines areas for further research. Illustrative examples drawn from the military training domain are used throughout the paper to describe key ideas.

2 An Ontology for Adaptive Simulation Based Training

The simulation based training method described in this paper seeks to address the training needs of defense missions. Mission requirements are the drivers for the training goals, which, in turn, drive the determination of warfighter training performance measures. An ontology (conceptual model) for adaptive simulation based training is shown in Figure 1. Knowledge, Skills, and Experiences (KSE's) must satisfy Mission Requirements as shown in Figure 1. 'Knowledge' is defined as "information or facts that can be accessed quickly under stress." Examples of knowledge areas include tactical plan coordination, team operating protocols, tactical maneuver principles, and team maneuver expectation templates [2]. 'Skill' is defined as "a compiled sequence of actions that can be carried out free of error under stress." Examples of skill areas include single mode selection to maximize information requirements, scan volume placement to maximize relevant information gathering, and radar control manipulation to locate and track relevant targets. An 'Experience' is defined by [3] as a "development event during training and/or career necessary to learn a knowledge or skill or practice a MEC under operational conditions." Dependencies between KSEs and Training Performance Measures provide an important requirement for determining the structure and content of the Performance Measures. Capturing these important dependencies is part of our strategy for designing mission-driven metrics as outlined later in this paper. In the context of simulation based training, Training Scenarios are decomposed into finer-grained building blocks called 'Drills.' A Drill is defined as

the smallest building block of training simulation content. Drills may be grouped together into meaningful collections for building Training Scenarios (Figure 1). A key idea is the notion of adaptively composing/assembling simulation training material from building blocks of re-usable parts: drills and drill collections [4].

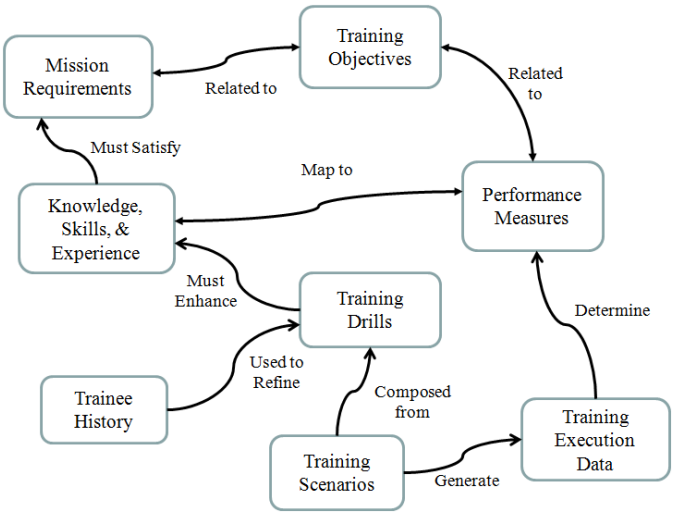


Fig. 1. Ontology Model for Adaptive Simulation Based Training

As indicated in Figure 1 the Training History of each student (often stored in 'Electronic Training Jackets') is used to individualize the training content (drills and scenarios) to address unique training gaps of different students. The execution of the scenario based training will generate simulation log data that is used to compute the values of carefully designed performance metrics as shown in Figure 1. The Performance Metrics must address Training Objectives and are mapped to the simulation drills. The Drills themselves are carefully engineered to induce the Skills and Experiences that address the warfighter mission requirements.

3 An Adaptive Simulation Based Training Method

This section describes a method for adaptive simulation based training. The method identifies the activities required for conducting simulation based training and the relationships between these activities in terms of the activity inputs, outputs, enabling mechanisms, and constraints. The IDEF0 function modeling method (www.idef.com) was used to represent the method, which is summarized pictorially in Figure 2.

The following paragraphs describe the adaptive simulation based training method in greater detail.

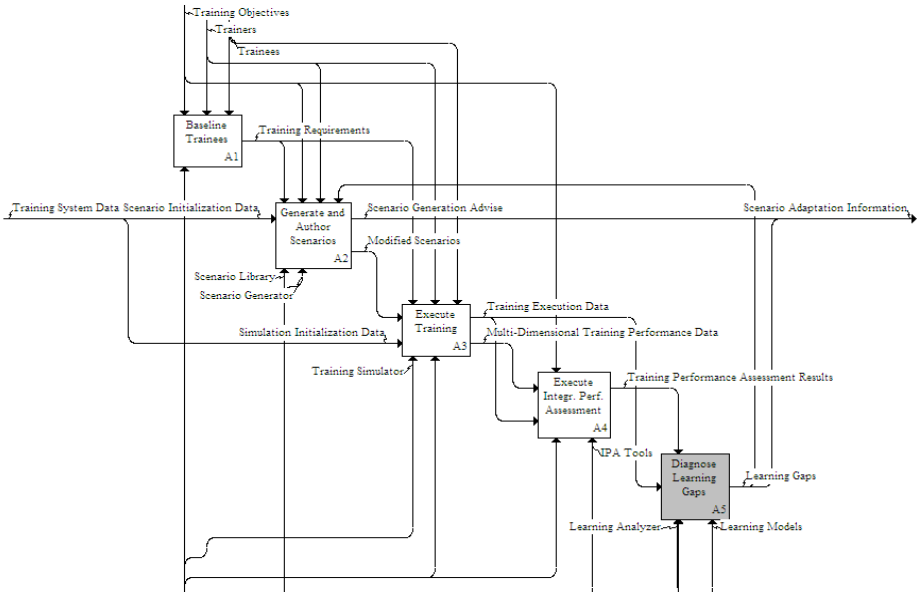


Fig. 2. Adaptive Simulation Based Training Method

Baseline Trainees: This activity involves performing benchmarking or testing performance at the point of entering the training sessions. For example, subjective evaluation methods may be used to baseline cognitive knowledge structures among trainees. The main outcome of this **activity** is a set of prioritized training requirements for the current set of trainees. These training requirements are used to inform subsequent activities in this methodology. For example, the training requirements are used to help determine the appropriate scenarios that best address the skill and experience needs of the trainees in the current training session. The baseline assessment results also provide a 'control data set' during the integrated training performance assessment activity.

Generate and Author Scenarios: This activity involves generating the following information: (i) identifying drills and collections of drills, and (ii) composing and sequencing drills to provide scenario design (and redesign) advice. A combination of rule-based methods and data analytics-driven methods may be used for scenario design information generation.

Execute Training: The scenario-based training simulation models are initialized with the mission-specific training data sets. The simulations are then executed. The training participants (trainees and instructors) interact with the simulation in a manner that induces the learning that is intended by the training objectives.

Perform Integrated Training Assessment: This activity involves using the results from the simulation based training event in order to measure the performance of the trainees (students). There are many different types and levels of performance metrics. The ability to scientifically measure performance is influenced by many factors including (a) the availability of a well conceived plan and the engineering design of

simulation training instrumentation, (b) the availability of data generated during the training session, and (c) the availability of subject matter experts (e.g., trainers and research scientists). In general, performance metrics span many dimensions. The metrics may be *quantitative/objective* or they may be *qualitative/subjective*. In some situations, metrics are *binary*; these metrics are used to determine whether or not a particular event/action was enacted by the trainee (a 'Yes or No' type metric). *Time based* metrics measure time intervals (e.g., actual time of a particular response vs. the desired response time) and *proficiency metrics* that determine the level of the correctness of a response (e.g., how well was the threat discriminated vs. a decoy, how well did the trainee respond to large amounts of noise and distracters in the data, etc.). The design of metrics is influenced by the types of warfighter missions and the degree of sophistication of the simulation based technology and infrastructure that is available for training. The design of sound metrics requires significant and intentional effort. More research is needed in some areas of measurement and metrics; for example, the evaluation of cognitive states often requires the use of sophisticated sensing technology such as neuro-physiological sensors. Our mission driven performance measurement approach involves (i) determining training performance based on objective training performance data, (ii) determining training performance based on subjective training performance data, and (iii) fusing the results of the different performance assessments to determine an aggregated assessment of training performance.

Diagnose Learning Gaps: This activity will infer the training gaps by comparing actual performance with desired performance.

4 Mission Driven Performance Measurement Approach

The overall strategy/rationale for training performance metric design is: The measures must provide a means to evaluate whether the warfighter is learning to be more effective in supporting missions. This implies that the training results must provide warfighters with the Knowledge, Skills, and Experiences (KSEs) needed to address mission requirements. These requirements are often met in different ways: basic training, mission qualification training, continuation training, etc. Simulation Based Training (vs. Classroom/Schoolhouse Training) is often used for Continuation Training (for refreshing and updating KSEs and addressing critical gaps that occur because of constantly changing mission requirements).

We now provide more details of our approach. A simple urban combat training example is used to illustrate the main ideas.

Step 1: Identify Knowledge and Skill (KS) Categories: This activity determines the set of knowledge and skills that are being imparted. We have identified multiple sets of KS categories for different types of warfighter missions based on the extensive body of knowledge that documents combat knowledge and skill sets (for example, [5-7]). For example, [5] lists Knowledge and Skill categories for building a clearing mission in Urban Operations (UO): (i) diagnosing and predicting, (ii) situation awareness, (iii) perceptual skills, (iv) improvising, (v) metacognition, (vi) recognizing anomalies, and (vii) compensating for equipment limitations. Our research indicates that these KS categories are inherently linked to decision processes (i.e., rationale/reasoning for making decisions and taking action).

Step 2: Identify Mission Specific Task Sets: This activity determines mission specific task categories that are relevant to live combat training. These tasks manifest themselves at multiple levels of granularity and specificity. To illustrate, we use a simulation based UO warfighter training situation. In this situation, there are two broad categories of decision requirements: *task-focused* and *task-independent*. Task-focused decision requirements for building clearing missions are (i) determine how to secure the perimeter, (ii) determine how to approach the building, (iii) determine how to enter the building, (iv) determine how to clear the building, (v) determine how to maintain and extend security, and (vi) determine how to evacuate the building. Likewise, the task-independent decision requirements for building clearing missions are (i) maintain the enemy's perspective, (ii) lead subordinates, (iii) maintain the big picture and situation awareness, (iv) project into the future, and (v) understand and apply rules of engagement. The decision requirements are linked to specific parameters found in tactics manuals that provide instructors with guidelines for measuring performance against recognized standards of employment doctrine. Each of the decision requirements (task-focused and task-independent) are governed by critical decision and judgments. To illustrate, the "determine how to secure the perimeter" decision requirement is linked/tied to the following critical decisions and judgments: (i) determining how to seal off the area, (ii) determining where to place security assets, (iii) determining which assets and people to employ, (iv) determining where to concentrate fire, (v) determining how to synchronize fire and the shifting of fire, and (vi) if multiple buildings are to be cleared, determining which to clear first. A simulation based UO training engagement will provide training events/drills that induce critical decisions and judgments for these decision requirements.

Step 3: Design Metrics for Knowledge/Skills and Tasks Combinations: This activity formulates performance metrics for meaningful associations of Knowledge/Skills with (mission specific) Tasks. Figure 3 illustrates, by example, the idea of a metric that is determined through the association of knowledge/skills with tasks. Mission-specific tasks are listed for a building clearing mission in UO operations. All the tasks within a mission are governed by Tactics, Techniques, and Procedures (TTPs) or Predeployment Training Program (PTP) standards. TTPs are composed of parameters that instructors monitor against recognized standards of employment doctrine. Each task is decomposed into individual actions that the soldiers within a unit or that a specific soldier should perform to successfully complete the task. Instrumented training facilities and instructors record measurements on actions performed by soldiers. Objective measures may be determined from the simulation output log data, and subjective measures are recorded by training instructors [usually Subject Matter Experts (SMEs)] using pre-determined grade sheets. The preferred choice of assessing team performance is through subjective measures.

Every mission requires units to possess certain Knowledge and Skills (KS) to successfully execute and accomplish the goals. These KS are applicable across all the tasks, but it is possible that the degree of association (weights) might vary significantly. For example, "Improvising" might play a significant role for the "Approach the Building" task than for the "Secure Perimeter" task. We envision, for each pair of

associations between KS areas and Tasks, that there exists metrics for quantifying team performance effectiveness. We have listed a few metrics and measures in the figure and how they relate to KS-Task association as follows:

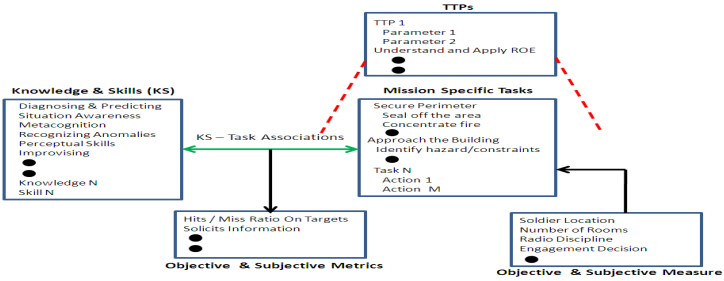


Fig. 3. Approach for Designing Metrics: An Example

- **Objective metric, "Hits to Miss Ratio On Targets":** this can be associated to the "Diagnosing and Predicting" and "Concentrate Fire" (*action within "Secure Perimeter" task*) pair.
- **Subjective metric, "Solicits Information":** this can be associated to the "Recognizing Anomalies" and "Identify Hazard/Constraints" (*action within "Approach the Building" task*) pair.

The above two metrics are given to illustrate the idea and approach for designing metrics.

Step 4: Determine Instrumentation Strategy: This activity refers to the creation of the means for deriving values of the performance evaluation metrics. This activity will be significantly influenced by the type of (post-) training data that is actually available within the simulation based training system.

5 Architecture of an Adaptive Simulation Based Training System

This section summarizes a conceptual architecture of a system that provides automation support for the 'Adaptive Simulation Based Training Method' described earlier in this paper (Figure 4). The functions supported by this architecture include (i) Integrated Training Performance Assessment, (ii) Learning Gap Diagnosis, and (iii) Adaptive Scenario Generation.

The architecture provides automated support for the adaptive generation of scenario based training simulations through (i) agile and comprehensive performance measurement and (ii) targeted training gap diagnosis. The Scenario Generation tool auto-generates scenario creation and scenario redesign information, allowing users to rapidly author/reconfigure scenarios to address the focused needs of the trainees. The scenario design advice is intelligently guided by the measured training gaps, the training objectives, and the desired performance goals. The Intelligent Performance

Assessment (IPA) Tools determine the values of training performance metrics by combining the results of three types of assessment: (i) neurophysiological-sensor based assessment, (ii) objective assessment, and (iii) subjective assessment. The IPA Tools use an 'information fusion' approach to integrate the measurements from the three different assessment methods. The Learning Analyzer compares the results of the IPA with the desired performance (based on the training objectives and the expertise level of the trainees). An important output of the Learning Analyzer is a prioritized set of trainee learning gaps that is addressable through redesign of the training scenarios. The architecture houses different types of knowledge models: (i) the Scenario Library, and (ii) the Knowledge Base (KB) that contains Rules, Fuzzy Rules, and Analytic Models.

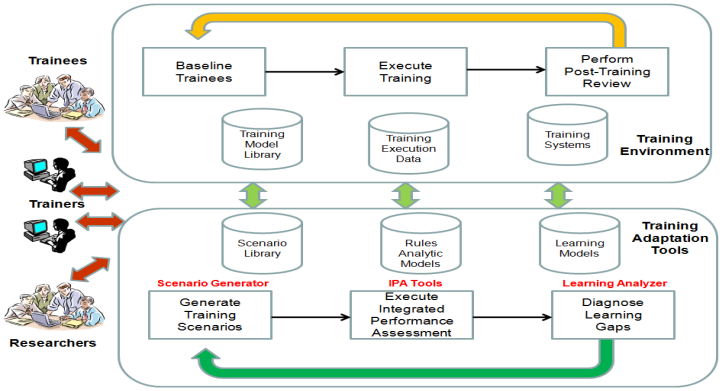


Fig. 4. Adaptive Simulation Based Training System Architecture

Finally, the Training Adaptation Tools subsystem interfaces with the 'Training Environment' that includes (i) Training Systems (this includes training simulators and the information infrastructure needed to manage the execution of simulation-based training), (ii) the Training Model Library (this refers to the collection of models used within the simulation-based training environment), and (iii) the Training Execution Data (this refers to the transactional data that is managed within the simulation-based training data; this includes simulation input data and simulation output data).

The adaptive simulation based training architecture has been implemented for air to air combat training with the U.S Air Force and the U.S. Navy. End user validation of this technology is currently ongoing within a laboratory setting [4].

6 Summary and Areas for Further Research

This paper described a structured method of adaptive simulation based training. Driven by an ontology model of adaptive simulation based training, the method characterizes the simulation based training activities and their interrelationships. A central element of the method is a mission-driven, information fusion-based approach for integrated performance measurement. Finally, an automation architecture is outlined

that provides a pathway for realizing the practical benefits of the adaptive simulation based training methods described in this paper.

Key benefits of the research described in the paper include (i) significant reductions in time and cost to develop and maintain simulation based training systems, (ii) improved effectiveness and quality of simulation based training in response to dynamically changing and complex training needs and requirements, and (iii) a component-based architecture that enables rapid, affordable, and scalable technology insertion and deployment.

Areas that would benefit from further research include (i) design of mission-driven metrics and instrumentation methods to measure intuitive decision making capabilities during simulation based training, (ii) design of methods to rapidly tailor simulation based training to address dynamically evolving mission-driven needs of individuals and teams, and (iii) design of hybrid training methods and tools that combine (a) simulation based training, (b) game based training, and (c) computer based training.

References

1. Cannon-Bowers, J., Salas, E. (eds.): Making Decisions Under Stress: Implications for Individual and Team Training, pp. 365–374. APA Press, Washington, DC (1998)
2. Alliger, Colegrove, Bennett: Mission Essential Competencies: New Method for Defining Operational Readiness and Training Requirements. In: Thirteenth International Occupational Analyst Workshop. San Antonio, TX (2003)
3. Rodriguez, D., Tossell, C., Garrity, M., Morley, R.: Promoting Air and Space Operations Center (AOC) Training Transformation by Quantifying and Refining AOC Training Scenarios. 2004 I/ITSEC Conference, Paper No. 1822 (2004)
4. Benjamin, P., Graul, M., Akella, K., Gohlke, J., Schreiber, B., Holt, L.: Towards a Method for Adaptive Scenario Management. In: Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Paper No. 12080 (2012)
5. Phillips, J., McCloskey, M.J., McDermott, P.L., Wiggins, S.L., Battaglia, D.A., Thordsen, M.L., Klein, G.: Decision-Centered MOUT Training for Small Unit Leaders. DASW01-99-C-0002 (2001)
6. Denning, T., France, M., Bell, J., Bennett, W., Symons, S.: Performance Assessment in Distributed Mission Operations: Mission Essential Competency Decomposition. Interservice/Industry Training, Simulation, and Education Conference, Paper No. 1616 (2004)
7. Lampton, D.R., Cohn, J.V., Endsley, M.R., Freeman, J., Gately, M.T., Martin, G.A.: Measuring Situation Awareness for Dismounted Infantry Squads. 2005 Interservice/Industry Training, Simulation, and Education Conference, Paper No. 2293 (2005)

Enhancing HMD-Based F-35 Training through Integration of Eye Tracking and Electroencephalography Technology

Meredith Carroll[1], Glenn Surpris[1], Shayna Strally[1], Matthew Archer[1], Frank Hannigan[1], Kelly Hale[1], and Wink Bennett[2]

[1] Design Interactive, Oviedo, Florida
{meredith,glenn.surpris,shayna.strally,matt.archer, frank.hannigan,kelly}@designinteractive.net
[2] Air Force Research Laboratory, Wright- Patterson Air Force Base, Ohio
Winston.Bennett@wpafb.af.mil

Abstract. The ever increasing complexity of knowledge, skills and abilities (KSAs) demanded of Department of Defense (DoD) personnel has created the need to develop tools to increase the efficiency and effectiveness of training. This is especially true for the F-35, the first 5th-generation aircraft to use an HMD as the primary instrument display. Additionally, the F-35 can perform operations previously performed by multiple operators, which potentially places incredible strain on the pilot's cognitive resources by exposing him to large amounts of data from disparate sources. It is critical to ensure training results in pilots learning optimal strategies for operating in this information rich environment. This paper discusses current efforts to develop and evaluate a performance monitoring and assessment system which integrates eye tracking and Electroencephalography (EEG) technology into an HMD enabled F-35 training environment to extend traditional behavioral metrics and better understand how a pilot interacts with data presented in the HMD.

Keywords: Training, Performance Assessment, Eye tracking, EEG, Helmet-mounted display, Heads-up display, F-35.

1 Introduction

According to the Air Force Transformation 2010, "the ultimate source of air and space combat capability resides in the men and women of the Air Force…first priority is ensuring they receive the precise education, training, and professional development necessary to provide a quality edge second to none"[1]. As technology progresses, the extensive knowledge, skills and abilities (KSAs) required of Department of Defense (DOD) personnel increases, and the demand for efficient yet effective training intensifies. This training need is particularly evident with 5th-generation tactical aircraft such as the F-35 Lightning II (formerly referred to as the Joint Strike Fighter).

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 21–30, 2013.

In the 1970's heads-up displays (HUDs) were introduced to tactical combat aircraft, which projected essential flight information onto the cockpit glass. This allowed pilots to continue to keep their eyes directed outside of the aircraft without being required to look down at important gauges. Several decades later, the development of the helmet-mounted display (HMD) allowed the HUD to be placed inside the pilot's helmet. The F-35 Lightning is the first 5th generation aircraft to use an HMD as the primary instrument and sensor display. Additionally, the F-35 is capable of performing air-to-air combat and air-to-ground strikes while flown by a single operator. Tactical information has been added to the HMD to aid the F-35 pilot in performing the additional tasks. This translates to an increase in cognitive demands for an F-35 pilot, with the amount of data from different sources potentially exceeding an individual's natural cognitive processing limits. As noted by Endsley (2001), data does not equal information, and may not be useful "unless it is successfully transmitted, absorbed and assimilated in a timely manner by the human" [2]. To date, the amount of information to be displayed often already exceeds available display space. Although essential information is provided on the HMD, the pilot must periodically transfer attention to other areas of the cockpit, such as the Multi-function Displays (MFDs), a paper copy of the Checklist, or cockpit control panels, in order to view more detailed information throughout a flight. Attentional demands in the cockpit shift frequently and rapidly if an emergency such as engine failure occurs, adding to the cognitive stress already amplified in an emergency situation. Given this, it is critical to ensure training results in pilots learning optimal strategies for operating in this information rich environment, including appropriate attention allocation between the different displays and pieces of information displayed within.

2 Training Needs / Opportunity

The application of HMD systems in tactical aircraft and simulation environments has substantial implications for performance assessment, proficiency tracking, and training. Much of the interaction that occurs with an HMD is unobservable, including gaze location/durations and cognitive processing of various information inputs that may not have an overt behavioral response. In order to effectively diagnose deficiencies/inefficiencies in performance and provide targeted feedback, it is necessary to obtain process level measures of performance that include capture of unobservable perceptual and cognitive tasks. To achieve this, there is a need for practical tools and instrumentation to better capture important data that can be assimilated in real-time to more accurately assess pilot performance, including data presented in the HMD, the interaction of the pilot with the data, and reactions and actions taken based on the data. With this enhanced data capture and performance monitoring capability, improved After Action Reviews (AARs) and debriefings will be possible that may substantially enhance training effectiveness and efficiency.

The current training practices for the F-35 lightning were investigated to ensure that the research effort to develop a precision performance assessment system, referred to as the Helmet-Mounted Display ASsessment System for the Evaluation of

eSsential Skills (HMD ASSESS), is designed to address the training needs of the F-35. The current training program for F-35 transition pilots is 8 weeks long. The transition pilots are comprised of legacy aircraft experts such as experienced F-16 or F-22 pilots. These pilots will become F-35 instructors upon the completion of the program. Training begins with a week of military lectures, followed by 3 weeks of lectures and academic courses specific to the F-35. A pilot training aid (PTA) laptop simulator is flown by transition pilots during these early phases of the course. The last phase of the training program is a mixture of 8-10 F-35 Full Mission Simulator (FMS) sessions and 4-5 actual flights in the F-35. The PTA and the FMS are the two main simulators used in the transition curriculum. The PTA has a large touchscreen monitor that displays both the out-the-window view of the aircraft as well as the touchscreen instrumentation (i.e. Main Forward Display). In addition to the touchscreen monitor, the PTA also has a full replication of the F-35 Hands-On Throttle and Stick (HOTAS). The PTA is mainly used during academic lectures to familiarize the pilot with the controls and procedures for the F-35. An HMD is not used in conjunction with the PTA.

The FMS is a high fidelity flight simulator which contains a full 1-to-1 replication of the F-35 cockpit surrounded by a dome with almost 360 degrees of visual coverage. The pilot trainee is outfitted with an HMD visor that reveals a HUD fixed on the center windscreen. Additionally, a de-cluttered, un-fixed version of the main HUD with a reduced selection of essential symbols (e.g., airspeed, altitude) appears on the HMD when the pilot turns his/her head off bore-sight (i.e., left, right, up, or down). The simulator sessions in the FMS are 1.5 hours in duration and are preceded by a 1 hour pre-brief and followed by a 1 hour debrief. Each trainee in the FMS has the individualized, one-on-one attention of an instructor. The instructor has an operator station where he can launch scenarios and insert abnormal aircraft conditions. During the training session, the instructor can also view the pilot's performance unfolding from a series of view, including the field of view (FOV) in the cockpit due to a head-tracker associated with the HMD.

The debrief then provides the opportunity for the instructor to playback any flight segment during the simulator session and review notes, exceptional performance, and trainee performance errors. Control inputs, the pilot's FOV, and other simulator information can be accessed by the instructor to facilitate this debrief. Instructors depend on overt behavioral actions and communications to identify performance errors. One limitation of this approach is the inability of the instructor to determine the specific instruments the pilot is monitoring, both within the HMD and on the MFD. Heads up/heads down status can typically be inferred based on the FOV presented by the HMD, however, the specific information that the pilot is visually integrating is not accessible. Given that a large portion of the task is monitoring information presented by a range of instruments; this limits the instructors understanding of how pilot performance is unfolding.

Without sufficient data collection and diagnosis of performance data, evaluations and feedback provided by instructors may not address the underlying sources of poor performance. There are multiple reasons for this, including: 1) instructors may not be able to detect all errors due to the high workload associated with monitoring a

complex scenario; and 2) instructors are unable to monitor subtle physical behaviors such as scanning patterns or attention allocation. As a result, instructors may not drill down far enough to expose the root cause of training deficiencies. For example, during irregular flight training such as warning or error procedures, a trainee may fail to take appropriate action to correct the aircraft parameters during a warning indicator. This could be due to several reasons including 1) he/she is not monitoring/scanning the relevant content in the cockpit, 2) he/she is monitoring the relevant content in the cockpit, but does not detect that they are out of tolerance, or 3) he/she detects they are out of tolerance but does not understand appropriate actions to take to mitigate. Additionally, there may be overarching error patterns undetected by the instructor, such as tendency to allocate unnecessary attention heads-down/ within the main forward display (MFD) or to symbols not relevant to the task at hand. Having data that can help instructors to determine the root cause of errors could provide key information regarding the general nature of the failures, which could potentially facilitate development of more effective training interventions.

Given the increased responsibilities and cognitive workload of the F-35 pilot, pilot's cognitive interactions with the HMD and other instrumentation are ever more important to ensuring that training feedback is as accurate and helpful as possible. To this end, objective measures of pilot information processing efficiency and effectiveness are required. Since information interaction within a head-mounted display (HMD) is limited almost entirely to perceptual and cognitive processes such as visual scan and information processing, there is a need for innovative solutions that can accurately and reliably capture this 'unobservable' behavior in order to 1) understand how an HMD is impacting pilot performance and 2) design training to effectively maximize performance. With the advancement in physiological monitoring technology such as eye tracking and EEG there is an opportunity to make these unobservable processes accessible to instructors to increase the accuracy and effectiveness of training feedback.

2.1 Eye Tracking and EEG

Visual attention can provide important insights to the information used in task performance, such as the importance of various features or cues [3]. Several studies [4; 5; 3; 6] have used eye tracking to extract information about scan strategies. These studies have demonstrated that eye tracking can aid in the assessment of perception through measurement of visual attention during observation via gaze, scan path, and fixation data. These measures can provide a means for increasing the granularity of performance feedback and hence the effectiveness of debriefs based on these measures. Additionally, mobile eye tracking technology has been successfully implemented in both the commercial flight deck [7] and military fighter jet [8] simulation environments to measure scan path sequence, visual attention allocation, overall situational awareness, and fixation times. These measures are particularly useful for assessing HMD interactions, as the HMD is an area of the cockpit where the pilot is solely monitoring information visually and is not performing observable direct control inputs. EEG has been successfully used in previous studies [9] along

with electrocardiogram (ECG) sensors [10] to measure trainee workload in the aircraft simulation environment. Cognitive workload is of particularly interest in the F-35 environment due to the previously-stated consolidation of duties. Such a measure could allow the identification of times when cognitive overload led to performance failures as opposed to skill decrements, allowing for feedback to more accurately target the root cause of errors.

Eye tracking and EEG measures have been successfully implemented together in a number of desktop-based environments to provide this deep diagnostic evaluation of performance [11, 12, 13, 14, 15]. Eye tracking and electroencephalography (EEG) can be used in combination to access such "unobservable" perceptual cognitive processes as scan strategies [11], attention allocation [12, 13, 14] and cognitive workload [15, 16]. Thus, eye tracking and EEG emerged as the most suitable combination of physiological measures to incorporate into HMD ASSESS to address the training needs of the F-35 Lightning II. The initial version of HMD ASSESS intended for use in the F-35 training environment will be limited to utilizing eye tracking measurements, with EEG measurements reserved for the version of HMD ASSESS used in conducting research. However, incorporating EEG in the training assessment version of HMD ASSESS is the end goal when EEG technology becomes more deployment friendly.

These measures can provide a means for increasing the granularity of performance feedback and hence the effectiveness of debriefs based on these measures. Specific advancements required to realize the benefit of such metrics in FMS include 1) integration of hardware into an HMD; 2) analysis techniques that can reliably identify visual focus, such as when focus is on the Heads-Up Display (HUD) versus out the window, on which instrument the pilot is fixating, and cognitive state (e.g., cognitive overload); and 3) display techniques for visualizing the data in a format usable by pilot instructors during assessment and debrief.

3 HMD ASSESS Approach

The HMD ASSESS development effort aims to create a precision performance assessment system which integrates advanced sensor technologies including eye tracking and EEG to measure "unobservable" perceptual and cognitive processes such as visual scan, attention allocation and cognitive workload during HMD-based performance. Based on these granular-level process measures, HMD ASSESS will diagnose performance deficiencies (e.g., failures in monitoring and detection) and inefficiencies (e.g., times of cognitive overload, distraction or inefficient scan strategies) and provide Real-time and After Action Review (AAR) summaries of individualized performance issues. These summaries can be used to 1) support training instructors in identifying skill decrements which need to be effectively remediated to achieve criterion performance and 2) assist system designers in gaining an understanding of how a pilot is interacting with the system and 3) identify specific problem areas within the display. HMD ASSESS will thus provide a comprehensive understanding of pilot performance within an HMD enabled environment, a task

previously unachievable due to the unobservable nature of these processes. The resultant precision performance assessment system is intended to improve training effectiveness by providing instructors with access to previously unobservable perceptual and cognitive processes, allowing them to pinpoint the root cause of performance deficiencies (e.g., issues with attention allocation) and effectively tailor the debrief to address the problem.

This effort commenced with the development of a taxonomy which delineated the data presented in the F-35 HMD and the expected pilot interactions with this information. F-35 instructor pilots and other domain experts were interviewed throughout the design process, with interviews conducted in an iterative manner. Utilizing the taxonomy as a foundation for what needs to be measured in order to understand pilot interactions with the HMD, a conceptual design of HMD ASSESS was developed and evaluated by F-35 Subject Matter Experts (SMEs) who provided input leading to the redesign of several HMD ASSESS metrics, diagnostic methods and displays. The resulting HMD ASSESS conceptual model consists of four main components, including 1) Measurement component, 2) Diagnosis component, and 3) Instructor Displays component.

These components are discussed in the following sections and a use case is presented to illustrate the tool concept of operations.

3.1 HMD ASSESS Conceptual Design

HMD ASSESS Measurement. The HMD ASSESS measurement and data capturing component will log the occurrence of relevant events during the training session. The measurement component will receive events from a variety of available data sources, including the simulation system or another instructor learning station as appropriate (e.g., when warnings are provided or HOTAS inputs received from the pilot), eye tracking (e.g., ocular fixations relative to pre-defined high or low priority areas of the cockpit for a specific segment of flight or emergency scenario), EEG hardware (e.g., cognitive workload levels) and input devices available to the user. The measurement component will assess events received for inclusion in the diagnostics to facilitate system flexibility required for integration into multiple training simulations. This component will be the hub for integrating the available data sources and calculating metrics to support the diagnostics.

Taxonomy Development. Based on an analysis of F-16 and F-35 operations, and advanced HMD systems (including the Helmet Mounted Display System, the Joint Helmet Mounted Cueing System, and the Helmet Mounted Integrated Targeting system), an HMD-ASSESS Taxonomy was developed to serve two purposes. First, it provides a preliminary understanding of unique and common data displayed across HMD systems as well as when and how pilots interact with HMD presented data. Second, it provides a foundation for identifying metrics to assess this interaction as it identifies when pilots should be monitoring different pieces of information and potential errors in doing so. The taxonomy provides a breakdown of the following information for 40 symbols provided in the F-35 HMD interface and 8 MFD displays, including a description and location of the symbol, other locations in the cockpit the

same information can be found, the associated tasks performed when interacting with the symbol, and common errors associated with monitoring the symbol.

HMD ASSESS Diagnosis. The diagnostic component will utilize the raw metrics output from the measurement component and run a series of algorithms utilizing constraint-based modeling approaches to identify key performance decrements and the underlying causes of these decrements such as insufficient attention allocation, cognitive state and occurrence of tunnel vision. This will be used to identify critical performance issues on which instructors should focus their training interventions such as AAR debrief and future training scenario selection and manipulation. Output of the performance diagnosis will provide instructors with pilot generated errors, root-cause error analysis, and consolidated error pattern analysis.

HMD ASSESS Real-Time Display. The HMD ASSESS will include a real-time presentation of pilot trainee eye scan data displayed over a video feed displaying the area of the cockpit where the trainee is currently looking. The real-time display will be viewable in the instructor station, so the instructor can monitor where the pilot is looking and flag errors if desired, in addition to the errors identified automatically by the system.

HMD ASSESS After Action Review Displays. An AAR screen generator will be implemented in HMD ASSESS that displays a variety of data to assist instructor in pilot performance assessment and debrief, including:

- Graphical representation of pilot eye scan performance
- Diagnostic information regarding pilot performance decrements
- Performance summaries of pilot behavioral and eye scan performance

Diagnostic outcomes will be fed forward to the display component which will present a single AAR screen containing 1) a playback mode showing real-time trainee eye scan data relative to pre-defined high or low priority areas of the cockpit for a specific training segment, 2) an overview mode showing a summary of all eye scan data relative to pre-defined high or low priority areas of the cockpit for a specific training segment, 3) a multi-level timeline which contains performance feedback and allows the instructor to zoom into specific segments of flight (i.e., taxi, take-off, approach, etc.), emergency scenarios (e.g., engine flame-out or rudder failure), instructor flags or system identified errors (e.g., when the pilot misses a required task for a specific segment of flight or emergency scenario), 4) a summary list of all instructor flags and system-identified errors, and 5) summaries of the total time the trainee fixated on different areas of the cockpit (e.g., the total time the pilot was heads up or heads down, the total time spent looking at HMD symbols or MFD pages). In summary, the AAR screens provide the instructor with the ability to select a specific segment of flight in order to review scan patterns, errors, and visual allocation timing information with the pilot trainee.

4 HMD ASSESS Use Case

HMD-ASSESS is designed to be utilized during the actual training session and debrief. A use-case was developed to demonstrate the HMD ASSESS concept of operations for F-35 FMS training sessions and is presented in summary in this section.

A typical training session in the FMS may include several abnormal malfunctions from which a pilot must attempt to recover. During this particular training session, the instructor has inserted an Integrated Power Package (IPP) failure into the scenario. As the pilot trainee attempts to recover from the IPP failure, he performs three key errors: 1) the pilot misses a critical checklist item (i.e., arming the backup oxygen system); 2) the pilot spends too much heads down time looking at his checklist and fails to scan his primary flight instruments (altitude, attitude, airspeed) at the necessary intervals; 3) the pilot develops tunnel vision on an area of the cockpit irrelevant to the appropriate task, e.g., determining the best place to land, resulting in a delay in conducting a critical checklist item (i.e., open RAM door).

After the training session in the simulator has ended, the instructor uses the HMD ASSESS after action review displays to facilitate his debrief to the pilot trainee as follows. The instructor is interested in assessing the students handling of the IPP failure, so the instructor clicks on this segment of the timeline and the timeline automatically zooms into the IPP failure event. The instructor points out overall timing summary for that segment to the pilot, including total time heads up vs. heads down and total time in high priority areas. The instructor can illustrate to the pilot trainee that he spent a large amount of time heads down while handling the IPP Failure.

The instructor then clicks on the first system identified error, which automatically zooms the timeline down to a system default of 30 seconds on either side of the error. The instructor plays back the error and points out that, based on the eye tracking data, the pilot was distracted from reading the checklist by focusing on blinking lights on the IPP Panel.

The instructor then moves on to the next error (i.e., breakdown in a periodic eye scan of flight instruments), by selecting the error from the error summary list. The instructor wants to show the pilot how he failed to scan his primary flight instruments frequently enough. By using the Overview mode containing a summary of all eye tracking data for 30 seconds on either side of the error, the instructor illustrates to the trainee that a scan of these three primary flight instruments did not occur during this time period. The instructor confirms this by pointing out the timing summary which shows that the pilot spent very few seconds looking at the altitude, attitude, and airspeed instruments for the specified window of time.

The instructor then points to the section of the timing summary that shows the total time spent on each MFD page for the segment of flight in focus. He uses this data to illustrate that the pilot spent only 30 seconds looking at the navigation page and flight instruments because he started to look for the nearest airport to land too early, instead of following the checklist steps. This caused the pilot to delay in opening the RAM (i.e., air intake) door, which resulted in systems overheating more quickly.

As illustrated in the se case, HMD ASSESS will allow an instructor to more accurately and efficiently diagnose a performance issue. Instructors will be better able to direct a pilot's attention during overwhelming flight scenarios and prevent pilots from making common mistakes with regard to visual attention allocation.

5 Future Research

Development of the HMD ASSESS prototype is currently underway. HMD ASSESS will be integrated with the PTA initially, with the ultimate goal of implementing the system in the F-35 FMS. As HMD ASSESS has an iterative lifecycle and development process, the initial HMD ASSESS prototype will be verified and validated, and then revised as needed following implementation. The effort will culminate with a training effectiveness evaluation to assess the impact HMD ASSESS has on performance assessment and training effectiveness.

Acknowledgements. This research was sponsored by the U.S. Air Force Research Laboratory under contract FA8650-12-C-6303.

References

1. United States Air Force. The Edge Air Force Transformation 2010 (2008), http://permanent.access.gpo.gov/lps40477/edgeweb.pdf (retrieved June 1, 2011)
2. Endsley, M.R.: Designing for Situation Awareness in Complex Systems. In: Proceedings of the 2nd International Workshop on Symbiosis of Humans, Artifacts and Environment, Kyoto, Japan (2001)
3. Raab, M., Johnson, J.G.: Expertise-based differences in search and option-generation strategies. Journal of Experimental Psychology: Applied 13(3), 158–170 (2007)
4. Jarodzka, H., Scheiter, K., Gerjets, P., van Gog, T.: In the eyes of the beholder: How experts and novices interpret dynamic stimuli. Learning and Instruction 20(2), 146–154 (2009)
5. Mello-Thoms, C., Ganott, M., Sumkin, J., Hakim, C., Britton, C., Wallace, L., Hardesty, L.: Different search patterns and similar decision outcomes: How can experts agree in the decisions they make when reading digital mammograms? In: Krupinski, E.A. (ed.) IWDM 2008. LNCS, vol. 5116, pp. 212–219. Springer, Heidelberg (2008)
6. White Jr., K.P., Hutson, T.L., Hutchinson, T.E.: Modeling human eye behavior during mammographic scanning: Preliminary results. IEEE Transactions on Systems, Man, & Cybernetics Part A: Systems & Humans 27(4), 494–505 (1997)
7. Weibel, N., Fouse, A., Emmenegger, C., Kimmich, S., Hutchins, E.: Let's look at the cockpit: exploring mobile eye-tracking for observational research on the flight deck. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 107–114. ACM (March 2012)
8. Wetzel, P.A., Anderson, G.M., Barelka, B.A.: Instructor use of eye position based feedback for pilot training. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 42(20), pp. 1388–1392. Sage Publications (October 1998)

9. Carroll, M., Fuchs, S., Hale, K., Dargue, B., Buck, B.: Advanced Training Evaluation System: Leveraging Neuro-physiological Measurement to Individualize Training. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2010(1), National Training Systems Association (January 2010)
10. Schnell, T., Hamel, N., Postnikov, A., Hoke, J., McLean III, A.L.: Physiological Based Simulator Fidelity Design Guidance (2012)
11. Carroll, M.B., Kokini, C., Moss, J.: Multi-Axis Performance Interpretation Tool (MAPIT) Training Effectiveness Evaluation Report (Program Interim Rep., Contract No. N00014-10-C-0091). Office of Naval Research, Arlington (2010a)
12. Hale, K.S., Fuchs, S., Axelsson, P., Baskin, A., Jones, D.: Determining gaze parameters to guide EEG/ERP evaluation of imagery analysis. In: Schmorrow, D.D., Nicholson, D.M., Drexler, J.M., Reeves, L.M. (eds.) Foundations of Augmented Cognition, 4th edn., pp. 33–40. Strategic Analysis Inc., Arlington (2007)
13. Hale, K.S., Fuchs, S., Axelsson, P., Berka, C., Cowell, A.J.: Using physiological measures to discriminate signal detection outcome during imagery analysis. Human Factors and Ergonomics Society Annual Meeting Proceedings 52(3), 182–186 (2008a)
14. Hale, K.S., Fuchs, S., Berka, C.: Driving EEG cognitive assessment using eye fixations. In: Applied Human Factors and Ergonomics 2nd International Conference, Las Vegas, NV, July 14-17 (2008b)
15. Carpenter, A., Fuchs, S., Whitlow, S., Fiore, S., Hale, K.: Auto-Diagnostic Adaptive Precision Training for Baggage Screeners (Screen-ADAPT). Phase I Final Report. Department of Homeland Security Contract No. N10PC20028 (2010); Carroll, M., Fuchs, S., Hale, K., Dargue, B., Buck, B.: Advanced training evaluation system: Leveraging Neuro-physiological measurement to individualize training. In: Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Annual Meeting, Orlando, FL (2010b)
16. Vision Systems International - JHMCS, `http://Vsi-hmcs.com` (retrieved September 15, 2010)

Bio-reckoning: Perceptual User Interface Design for Military Training

Tami Griffith[1], Deanna Rumble[2], Pankaj Mahajan[2], and Cali M. Fidopiastis[2]

[1] U.S. Army Research Laboratory, Human Research and Engineering Directorate, Simulation and Training Technology Center, Orlando, FL USA
Tami.Griffith@us.army.mil
[2] University of Alabama at Birmingham, School of Health Professions, Birmingham, AL, USA
{ddrumble,pmahajan,cfidopia}@uab.edu

Abstract. Simulation based training is one way to attain operational realism for training complex military tasks in a safe, task relevant manner. For successful transfer of knowledge, skills, and abilities to the dynamically changing military environment, the human-computer interface should minimally support learning during the training process and provide congruent action plans that facilitate understanding of the overall training goal. While there are emerging controller technologies, simulators still rely on such input devices as mouse and keyboard. These devices potentially cause information and training bottlenecks as they limit naturalistic interactivity within the more advanced serious gaming platforms. Given the shortcomings of current interface design, we suggest a human-computer interface framework that includes perceptual user interface components and an open source serious game testbed. We discuss a multimodal framework called bio-reckoning that integrates brain-computer interface techniques, eye tracking, and facial recognition within EDGE, the U.S. Army's newest serious game based training tool.

Keywords: simulation based training, perceptual user interfaces, brain-computer interfaces, serious games, military training, augmented cognition.

1 Introduction

Human Computer Interface (HCI) techniques do not enjoy the same timely advances as do computer components and related hardware. This lag is apparent when reviewing interface design for military training simulations, especially those following a serious game platform [1]. Smith's review of the use of games in military training makes clear that throughout history the game play or simulation supported cognitive function and related action needed for battlefield success. However, regardless of the technological advances (e.g., high fidelity terrain maps and realistic avatars) in today's serious game training paradigms current HCIs execute our 'plans to perform' an action in computer space by means of intermediary physical manipulations such as pressing keys or directing a joystick. Transferring actions through these traditional

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 31–40, 2013.

input devices places an intermediary between the human operator and the training simulation that can detract from training goals and objectives, and more importantly fail to support transfer of training to the field environment [2].

For example, computer-aided training that relies on joystick manipulations may hinder necessary cognitive processes (e.g., focused attention) through the bottleneck of translating intended action through an unnatural modality. Further, mapping computer interactions through these peripherals requires time to learn and to operate pretraining, while demanding time to translate a user's physical action into a limited predefined set of object behaviors during training. Given the artificiality of these input devices, physically operating interfaces for multiple objects may demand more cognitive resources and may lead to overload simply through motor control processing and motoric interference.

Also neglected in the traditional user-interface paradigm is the affective training that is necessary to provide the correct amount of emotion regulation required to react appropriately under stress [3]. The gaming industry leads the software and hardware development for many serious games for military use [4] [5]. Matching interactivity and emotional regulation design elements to the training environment is not important. The gaming industry's goal is entertainment, and not transfer of training to the theatre of war. Poor transfer of training from simulation to field is not only costly from a financial perspective, but also can lead to loss of life. Offering a more naturalistic interaction within military training simulations is an overlooked necessity.

How to proceed in creating appropriate user interfaces for supporting transfer of training is a non-trivial task. Defining types of actions necessary for task training and mapping them to the serious game action codes is a primary concern. Once actions are chosen, how to instantiate these codes in the serious game environment by choosing or developing serious game controllers is key. With new gaming technologies such as the Microsoft Kinect, a motion based controller, one strategy is to gather all state-of-the art controllers and user test for ease of use and improved performance. Problems quickly arise in that these controllers work optimally with a particular gaming console. Even controllers with their own software development kits pose interfacing issues that may require knowledge from a highly trained technician to integrate.

Another issue when choosing off-the-shelf gaming environments for training is that the metaphor used for interface design may not match the one needed for training. Controller technology and subsequent action code responses are gaming specific and serve the purpose of gaming goal, of which high entertainment value is one. These elements are also chosen as part of the gamming narrative or story that provides the nature of the interaction as a gaming element. How challenging these action pairings to gaming objectives are to learn depends on the overall goal of the game. For example, discovering how to launch a weapon may be a gaming objective for an action game. Action codes within the game environment support this aspect of exploration. In contrast, guessing how to change mission critical entities within military simulations for training is never appropriate. More importantly, the inability to access the source code of proprietary serious game platforms does not allow changes to controller parameters and their associated action maps further limiting the number, variety, and type of controller.

Low Cost Game Interfaces. Nintendo opened the game industry to non-gamers by creating an interface, Wii Mote, that made playing games more natural and engaging [6]. Researchers [7] explored the use of the Wii Mote and Numchuk for navigation, object manipulation, and object selection within a First Person Shooter (FPS) type gaming environment. The user navigated using the Wii Mote; the combination of the Wii Mote and the Numchuk performed the action of selecting and manipulating objects. Experienced FPS gamers reported the navigation strategy as frustrating; yet, many found the manipulation tasks more pleasant using the Wii tools as compared to a mouse. In this example, user satisfaction could be due to the optimized interactivity when using the new controllers. This research demonstrates the ease of use of these controllers for manual tasks, but falls short when describing the entire user experience.

Microsoft™ responded to the market success of the Nintendo Wii with the groundbreaking Kinect depth sensor camera [8]. The Kinect uses an infrared laser projector combined with an image sensor, which captures video data in 3D [9]. It is capable of simultaneously tracking six people with two active players at a time, allowing facial feature extraction or the ability to "recognize" players and the ability to track 20 joints per player [10]. The system includes a directional microphone to support voice control. In June of 2011, Microsoft™ released a non-commercial Software Development Kit (SDK) for use with Windows [11]. The Kinect interfaces with a standard PC via USB connector. This system has arguably changed the face of the interface world by bringing the player into the game more accurately than ever, however there are still unsolved problems that with the Kinect. Though the microphone is useful for administrative functions, it is still not reliable enough to replace the keyboard (or in this case, joystick). Gestures used to do interactions or administrative functions can be awkward or might not be readily recognized [12].

Voice as a Controller. Voice recognition for use in games is still at an early state of research. The goal of using voice in a FPS game is to assist in interacting with objects. For example, if the user approaches a vehicle, an action menu appears asking if the user wants to enter the vehicle. To activate the menu the player would use a specific word. Drawbacks to using voice as an interface are that background noise and casual conversation may unintentionally activate a task. It may be necessary to confirm direction to avoid false positive responses. The proposed strategy for this research would be to use a Small-Vocabulary/Many User system, with only a small set of words in use at specific times. Mohanram [13] showed that speech recognition as a game interface was not yet ready for public adoption with only 40% of the users considering speech as a better input strategy than voice. The greatest issue was misrecognition of the voice cues.

Apple has since released Siri as an alternative means of inputting data into its smart phone. Siri “understands” conversational context and is surprisingly accurate in converting spoken word into text [14]. Siri and PC applications (e.g., Dragon Dictation) demonstrate that voice recognition as a natural and intuitive interface tool is beginning to ‘come of age’. This technology clearly brings new functionality that was not readily available in the past.

Controller Testbed. A more efficacious serious game design strategy would be to start with a framework that provides a more systematic manner of testing controllers, interfaces, and content for their effects on military training and training transfer. The serious game platform should allow for full access to technology and action codes, as well as allow for the integration of multiple action controllers. This work uses Enhanced Dynamic Geo-Social Environment (EDGE), a military relevant gaming environment, to provide a testbed from which to assess gaming and training elements. Within EDGE, the Bio-Reckoning Interface (BRI) integrates multiple psychophysiological and body (e.g., facial features and limb movement) measures and uses these measures as naturalistic input control to the EDGE platform. For the purpose of this paper, we discuss the development of a BRI interface that uses brain computer-interface techniques, eye tracking, and face recognition to provide action codes to EDGE.

2 Perceptual User Interfaces for Military Training

Perceptual User Interface (PUI) design takes into account naturalistic human nonverbal and verbal human responses as part of the device or sensor input to the human-computer system [15]. This relationship between the trainee and simulator is symbiotic, like that in an intelligent automated system. This idea extends the concepts of Augmented Cognition, where the system uses the trainee's psychophysiological data to determine learner biophysical states that impede the learning process during training [16]. However, unlike the Augmented Cognition closed-loop system, the multimodal interaction capabilities of a PUI based system would allow for user control over the type of interaction that accounts for individual differences in how a person processes complex cues and related action in operational environments.

PUIs integrate concepts from perceptive, multimodal, and multimedia user interface designs. According to [15]: 1) perceptive interfaces are aware of the learner's body, face, and hands; 2) multimodal interfaces use several learner perceptual modalities such as speech and eye tracking as system input; and 3) multimedia systems include the use of text, graphics, animation, voice, and touch to best deliver training content. For these interface styles to be effective they must reciprocally monitor user behavior, model the goals and objectives of the training, flexibly change to learner preferences, positively support learning acquisition, support multi-tasking, and motivate the learner to interact meaningfully with to-be-learned material. Turk [17] contended that an ideal user interface should seamlessly transfer the intent of the user to the system, and the system response should appropriately support the user experience (e.g., reduce extraneous cognitive load).

2.1 PUI Military Examples

QuickSet is one of the first examples of a military based PUI that was a wireless, handheld capability that could control distributed interactive simulations based on

Modular Semi-automated Forces representing training at 29 Palms, California [18]. QuickSet used multiple input sensors such as speech, gesture, and direct object manipulation to support platoon leaders and company commanders with decision making involving multiple distributed assets (e.g., vehicles or personnel). The use of data fusion algorithms such maximum likelihood estimators coupled with artificial neural networks assisted to disambiguate the sensor inputs and increase reliability in noisy military exercises [19]. The system was also extensible to support 3-D terrain visualization.

Improved sensor technology and data fusion algorithms allowed for the miniaturization of sensors such that wireless, unobtrusive biosensors fit into wearable systems that convey real-time data acquisition [20]. The Virtual Locomotion Controller is an example of a wearable multimodal capability that used solid-state gyros and accelerometers, ultrasonic range sensors, and force sensitive footpads to provide naturalistic motion (e.g., crouching and running) within Military Operations on Urbanized Terrain (MOUT) simulated environments [21]. These advances in sensor and algorithm development also allow for redundant HCIs that present a combination of system features (e.g., face recognition, eye tracking, and graphics) to the operator, as well as user selection modes where the operator chooses the type of feature based on task relevance.

2.2 Bio-Reckoning Interface Components

The BRI is an example of a state-of-the-art PUI that provides data fusion across a multimodal sensor suite. These data streams are synchronized and applied to the serious game either as emulators for controllers in the case of proprietary games or as actual controllers as in training simulations. Their output can also provide information on the efficacy of the training system (i.e., eye tracking). Within this BRI prototype, we explored the use of BCI techniques, eye tracking, and face recognition.

Brain-Computer Interface Techniques. Brain-Computer Interfaces (BCIs) afford the possibility of removing the interface-as-middleman in both gaming and virtual reality contexts [22]. A typical BCI system consists of three processing modules: 1) a brain activity-monitoring device (i.e., electroencephalography-EEG) that records brain activity, 2) a signal-processing module that identifies specific brain patterns or features related to a person's intention to initiate action, 3) and a translator that converts these brain features into meaningful control commands [23]. Electrophysiological sources of control (ESC) are the mental activities and their associated EEG measures that become the control mechanism that perform actions within a given application. ESC are currently elicited in an active (user conscious control without external stimulation), a reactive (external stimuli elicits user brain response), or a passive (brain activity associate with a cognitive state drives system change) manner. The proposed BRI system combines an easy-to-apply wireless EEG sensor headset made by Advanced Brain Monitoring.

Eye Tracking. Blinks, direction of gaze, and fixations are all candidate eye movements that can act as naturalistic input to a serious game [24]. Additionally, gaze patterns provide information on how naturalistic a task appears to the operator [25]. These gaze patterns can also address training design elements through the evaluation of fixation patterns during task performance. For this work, we used an EyeTech TM3 eye tracker that monitors head movement as well as the gaze pattern of both eyes.

Face Recognition. Humans have a biologically mediated expertise in identifying faces versus objects [26]. Researchers in computer vision have sought to replicate this process in computer software since the 1960's. Facial recognition software measures various generalizable features common to all human faces (e.g., spacing between the eyes). While facial recognition systems are becoming more accurate at their primary task of identifying faces, techniques for transferring these facial features and their related meaning (e.g. smile) to an avatar is not readily available. Open source solutions exist; however, we chose a more robust commercial product, faceAPI created by Seeing Machines.

3 Enhanced Dynamic Geo-Social Environment (EDGE)

Experiential learning is one of the benefits of using a serious gaming platform for training military tasks. However, there are challenges to serious game use in training. Besides adoption, there needs to be a clear training benefit to using this training paradigm. Additionally, PUI features should serve to augment training or otherwise not be a part of the training system. Proprietary games for training do not allow code access to develop and test appropriate training content integrated with candidate PUIs. A solution that 1) provides access to source code; 2) allows for community input assisting with extensibility and updates; and 3) leverages coding expertise from a global network [27] affords the opportunity to test different types of PUI features within an operationally relevant training environment.

Fig. 1. Screen shots from EDGE showing accurate physics, terrain, and visual representations of military relevant operational environments

EDGE is a government owned architecture designed using AMSAA approved standards (e.g., *OneSAF*) to provide highly accurate virtual simulations of military operational environments utilizing state-of-the-art Multiplayer Online Gaming

(MOG) technologies [28]. Access to EDGE requires Federal Government sponsorship for use. This feature allows the community to upgrade the software to meet the challenge and pace of changing technology. Figure 1 depicts the fidelity of the computer-generated models, along with accurate physics, and military relevant operational environments.

The first level developed within EDGE was a tutorial level that requires the user to walk and run in each direction, complete a high and low crawl through small openings, walk a balance beam, drive a vehicle and shoot a weapon. Throughout the level, smart menus appear to interact with objects, such as entering a vehicle. If the user is at a keyboard, the "F" key activates these menus. This level ensures that users are familiar with controls prior to using a training level. However, for this research, the level ensures that the user can complete tasks within a FPS game environment. Additional levels allow free exploration of a small village and an urban environment.

4 Future Work Additional Controllers Integrated with the BRI

The next phase of BRI development will include the Playstation Kinect sensor, which will extract body positioning information and collect voice data. In addition, a Nintendo Wii will have a dual purpose of controlling a tactical weapon to allow the user to engage an enemy and a steering wheel to allow the user to have a sense of driving a vehicle. This combination will allow the user to move forward or backward using the BCI, jump, kneel, turn using the Kinect, shoot and drive using the Wii, and interact with the user prompts using voice. While the implementation of these controllers is initially to improve interaction within the EDGE platform, the intention is that the interface is not platform specific. Future research includes a phased approach described below.

4.1 Phase I: Establish Prototype and Measure User Experience

Implement a Prototype Interface Set to Engage with a First-Person-Shooter Environment. A difficult challenge is ensuring that all integrated controllers function in a complimentary manner. For example, if a player intends to jump across a hole in the training simulation, the Microsoft™ Kinect should detect that the player is jumping *and* the BCI must move the avatar forward concurrently. Otherwise, the player will not effectively complete the jump across the opening. Additionally, a single Graphical User Interface (GUI) should setup pairings for the controllers and their associated actions, as well as monitor pairing functionality throughout the experience.

Demonstrate the Prototype and Establish Measures of User Engagement. The first study will compare traditional keyboard and mouse input to the experience of using the BRI controllers. Measures of user experience will include: 1) time to successfully complete the tutorial level, 2) usability of the combined interface during interaction in a simulated small town, and, 3) psychophysiological measures of

engagement and distraction (e.g., brain activity and skin conductance). Finally, participants will respond to the NASA TLX and a user questionnaire based on five criteria for user acceptance of the interface (i.e., intuitive, readily available, augments existing user capabilities, accessible through an open toolkit, and fun).

4.2 Phase II: Comparison between Virtual and Live Experiences

Create a Scenario in the Virtual Environment That Replicates a Live Military Training Scenario. A MOUTexercise translatable into a live experience may provide the testing environment for this phase of research. A training exercise of this type would take place in a small simulated or mock-up town. The live exercise would use laser training weapons rather than live-fire weapons to reduce the risk of injury. The target location for the scenario is at the Maneuver Center of Excellence at Fort Benning, Georgia.

Compare Brain Activity and Skin Conductance Readings. To determine the how the level of realism experienced in a FPS game with an immersive interface suite compares to a live training experience at the same level of physical risk, trainees would experience both environments and perform similar tasks. Brain activity and skin conductance provide comparison measures of engagement and distraction or stress in each environment.

Compare Performance in the Live Environment after Practice in the Virtual Environment. Simulations historically reduce the cost and risk associated with live training. Further, simulation based training may better prepare a trainee for live training and ultimately for operational engagements. This study will compare the performance of trainees at the Maneuver Center of Excellence live training environment with and without preparatory virtual training. The expectation is that Trainees with virtual training preparation will perform better in a MOUT operation (building clearing, hostage rescue, etc) than those moving directly into a live training environment. This fits well with the described research because the level of realism established while in the virtual environment may be a critical factor in live training preparedness. To further establish that realism is a factor, three study groups will be established; one with no virtual training, one with keyboard and mouse at a desktop and one using the prototype BRI.

5 Conclusion

The ultimate goal of this research is to show that a combination of off-the-shelf emerging controller technologies integrated within a simulation-based trainer can improve the interaction between the human and computer. This improved interface can increase the user's sense of presence, immersion and flow, which may lead to improved human performance [29] and potentially to training realism. By creating a

PUI testbed using a military relevant simulation based training environment, this could benefit the gaming world as well as support military training applications.

It is clear that there is no one-size-fits-all interface in existence today. The premise of this research is that a combination of interface tools may begin to close the gap between the user and the immersive environment. Various modalities are mixed and matched and can be adjusted to support specific training needs; however, this type of experimentation should occur in a valid testing environment. In most cases, a traditional interface is sufficient; however, when total immersion is the goal for training or even for entertainment, a combination of interfaces may provide a better user experience.

References

1. Smith, R.: The long history of gaming in military training. Simulat Gaming 41(1), 6–19 (2010)
2. Grant, S.T., Barnett, J.S.: Evaluation of wearable simulation interface for military training. Hum. Fact. (2012), doi:10.1177/0018720812466892
3. Rahman, M., Balakrishanan, G., Bergin, T.: Designing human–machine interfaces for naturalistic perceptions,decisions and actions occurring in emergency situations. Theoretical Issues in Ergonomics Science 13(3), 358–379 (2012)
4. Smith, R.: Game Impact Theory: The Five Forces That Are Driving the Adoption of Game Technologies within Multiple Established Industries. Games and Society Yearbook, 1–32 (2006)
5. Smith, R.: The Disruptive Potential of Game Technologies. Research Technology Management 50(2), 57–64 (2007)
6. Klochek, C., MacKenzie, I.S.: Performance measures of game controllers in a three-dimensional environment. In: Proceedings of Graphics Interface 2006, pp. 73–79. CIPS, Toronto (2006)
7. Fischer, L., Oliveira, G., Osmari, D., Nedel, L.: Finding Hidden Objects in Large 3D Environments: the Supermarket Problem. In: Proceedings of 2011 XIII Symposium on Virtual Reality, pp. 79–88. IEEE Press, Brazil (2011)
8. Thorpe, A., Ma, M., Oikonomou, A.: History and Alternative Game Input Methods. In: The proceedings of the 2011 16th International Conference on Computer Games (CGAMES), Derby, UK, pp. 76–93 (2011)
9. Xia, L., Chen, C.C., Aggerwal, J.K.: Human Detection using depth information by Kinect. In: The proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 15–22. IEEE (2011)
10. Joystiq, `http://www.joystiq.com/2010/06/19/kinect-how-it-works-from-the-company-behind-the-tech`
11. Microsoft, `http://www.microsoft.com/en-us/kinectforwindows/`
12. Totilo, S.: `http://kotaku.com/5680501/review-kinect?skyline=true&s=i`
13. Mohanram, N.K.: A Speech-based Quiz Game. Final Dissertation Report (2003)
14. Sadun, E., Sande, S.: Talking to Siri: Learning the Language of Apple's Intelligent Assistant. Que Publishing, USA (2012)
15. Turk, M., Robertson, G.: Perceptual user interfaces. Communications of the ACM 43(3), 33–34 (2000)

16. Nicholson, D.M., Fidopiastis, C.M., Davis, L.D., Schmorrow, D.D., Stanney, K.M.: An adaptive instructional architecture for training and education. In: Schmorrow, D.D., Reeves, L.M. (eds.) Augmented Cognition, HCII 2007. LNCS (LNAI), vol. 4565, pp. 380–384. Springer, Heidelberg (2007)
17. Turk, M.: Perceptual User Interfaces. In: NSF Workshop (2006)
18. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J.: QuickSet: Multimodal Interaction for Simulation Set-up and Control. In: The proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC, USA (March 1997)
19. Pittman, J., Smith, I., Cohen, P.R., Oviatt, S.L., Yang, T.C.: QuickSet: A multimodal interface for militarysimulation. In: The Proceedings of the Sixth Conference on Computer Generated Forces and Behavioral Representation, pp. 217–224. Univ. of Central Florida, Orlando (1996)
20. Fidopiastis, C.M., Wiederhold, M.: Mindscape Retuning and Brain Reorganization with Hybrid Universes: The Future of Virtual Rehabilitation. In: Schmorrow, D., Cohn, J., Nicholson, D. (eds.) The PSI Handbook of Virtual Environments for Training & Education: Developments for the Military and Beyond, vol. 3, pp. 427–434. Praeger Security International, Westport (2008)
21. Lane, S.H., Marshall, H., Roberts, T.: Control interface for driving interactive characters in immersive virtual environments Technical Report, US Army Research (2006)
22. Lalor, E.C., Kelly, S.P., Finucane, C., Burke, R., Smith, R., Reilly, R.B., McDarby, G.: Steady-state VEP-based brain-computer interface control in an immersive 3D gaming environment. Eurasip J. on Appl. Sign Process. 19, 3156–3164 (2005)
23. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtschellere, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Clin. Neurophysiol. 113(6), 767–791 (2002)
24. Smith, J.D., Graham, T.C.N.: Use of eye movements for video game control. In: The proceedings of the ACE 2006, Hollywood, California, USA, June 14-16 (2006)
25. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends Cogn. Sci. 9(4), 188–194 (2005)
26. Tanaka, J., Gautghier, I.: Expertise in object and face recognition. The Psychology of Learning and Motivation 36, 83–125 (1997)
27. Darken, R., McDowell, P., Murphy, C.: Open Source Game Engines: Disruptive Technologies in Training and Education. In: Proceedings of theInterservice/Industry Training, Simulation and Education Conference (I/ITSEC). National Defense Industrial Association, Orlando (2005)
28. Dwyer, T., Griffith, T., Maxwell, D.: Rapid Simulation Development Using a Game Engine-Enhanced Dynamic Geo-Social Environment. In: The proceedings of Interservice/Industry Training, Simulation & Education Conference (I/ITSEC). NTSA, Orlando (2011)
29. Eisenberger, R., Jones, J.R., Stinglhamber, F., Shanock, L., Randall, A.T.: Flow Experiences at work; for high need achievers alone? J. Organ Behav. 26, 755–775 (2005)

Taiwanese EFLs' Metacognitive Awareness of Reading Strategy and Reading Comprehension

Yen-ju Hou

Shu Zen College of Medicine and Management, Taiwan
yunju@ms.szmc.edu.tw

Abstract. The study aims to identify the types of metacognitive awareness of reading strategies that Taiwanese EFLs (English as Foreign Language) used at medical junior colleges. In addition, metacognitive awareness of reading strategies were investigated to discover whether or not it affects students' English reading performance, specifically in reading comprehension. A total of 454 junior college students participated in the study. The results indicated that problem-solving reading strategies were used the most, followed by globe reading strategies, whereas support reading strategies were used the least. Regarding of the effects of variables on English reading performance, overall reading strategy use, and problem-solving reading strategies each significantly predicted students' reading comprehension. It's hoped that the finding could be helpful for further study as well as teaching.

Keywords: Metacognitive awareness, reading strategy, reading comprehension.

1 Background

Traditional instruction for reading often stresses on the teaching of vocabulary and grammar. Students are required to spend most of the time memorizing words, structures and grammar. Moreover, it fails to promote students' comprehension but results in their fear and rejection toward reading. During the studying period, the ultimate goal of teachers and students is how to come up with correct answers on exams. Thus, personal opinion and thoughts do not receive lots of attention on students' reading process, and the opportunity of using strategies to enhance reading comprehension is neglected.

A key element to enhance comprehension is metacognition, which is a personal's awareness to manage and monitor the process of cognition. Metacognitive strategic knowledge is a thinking ability involved in the process of reading comprehension. Moreover, students are able to adopt these strategies while reading so as to promote their reading comprehension (Baker & Brown, 1984; Yang, 2002).

In order to investigate the relationship between metacognitve reading strategies and comprehension, the study aims to identify the types of metacognitive awareness of reading strategies that Taiwanese EFLs (English as Foreign Language) used at medical junior colleges. In addition, metacognitive awareness of reading strategies were

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 41–49, 2013.

investigated to discover whether or not they affect students' English reading performance, specifically in reading comprehension.

In light of intense teaching schedule, it failed to provide students enough time and chance to notice their use of language learning strategies. Thus, it is hoping to offer students and teachers information about students' reading strategies, and to design suitable curriculum for students to cultivate their reading strategies and promote reading comprehension.

2 Literature Reviews

Reading comprehension has been identified as the cognitive skill that people use to comprehend what they read. Although reading in a native language (L1) is not the same as reading in a second language (L2), reading in L1 and L2 still share a similar process that can be influenced by various factors in different patterns (Cook, 2001).

Each student who enters the classroom comes from different family background and possesses different learning style. Several characteristics are synthesized to help obtain concepts about so-called good reader, and that are: (1) be active and positive reader; (2) know how and when to use different strategies in order to help them comprehend what they are reading; (3) tend to make assumption of any unclear or unfamiliar part on reading; (4) monitor how much oneself comprehend the reading; (5) manipulate different strategies to promote comprehension, and adjust strategies to compromise the part which failed to comprehend successfully by using previous strategies; (6) possess linguistic awareness (Celce-Murcia, 2001; Houtveen and Van de Grift, 2006; Maria, 1990 ; Tompkins, 2005).

Although there are various factors that affect the ability of reading comprehension, the key factor is strongly associated with the reader' cognitive skills and metacognition, such as if the reader has enough prior knowledge to link what they have learnt to the new information, using different strategies to help comprehend reading, and so on (Celce-Murcia, 2001; Farley and Elmore,1992). Besides, Flavell (1979) stated that metacognition is a personal's awareness to manage and monitor the process of cognition. Metacognitive strategic knowledge is a thinking ability involved in the process of reading comprehension. Moreover, students are able to adopt these strategies while reading so as to promote their reading comprehension (Baker & Brown, 1984; Yang, 2002).

3 Methodology

3.1 Participants and Population

Participants in this study mainly consisted of students who were learning English as a foreign language. Except for those in the foreign language department, students are required to take English courses for three hours per week their first three school years.

Before attaining junior college status, all students had been taking English courses for at least three years in junior high school.

The data for this study was based on surveys given to the students from different departments. These students contained different levels from basic to advanced English proficiency for the population of approximately 550 students. By removing the uncompleted surveys, valid samples were reduced to a total of 454 full-time students, shown in Table 1, including 100 males (22%) and 354 females (78%).

In addition to gender, participants are mainly from the following department: Applied English (34.1%), Nursing (31.5%), Physical Therapy (15.4%), Dental Laboratory Technology (11.5%), and Occupational Therapy (7.5%). The majority of participants were first-year junior college students (75.1%) at the age of 16 to 17 years old, whereas the rest students are second- to fourth-year students (24.9%). In addition, Chinese was their native language.

3.2 Research Instrument

The research instruments in the study were surveys that included two sections: Metacognitive Awareness of Reading Strategy Inventory (MARSI, Mokhtari and Reichard, 2002), and a reading test called the General English Proficiency Test (GEPT). As for the period used for answering, students were able to complete all within 90 minutes.

Section 1: MARSI. The MARSI, developed by Mokhtari and Reichard (2002), was used to identify 6^{th}- 12^{th} grade students' awareness and perceived use of reading strategies while reading academic or relative materialss. It is composed of 30 items in 3 scales: Globe Reading Strategies, Problem-solving Reading Strategies and Support Reading Strategies. In order to reduce difficulty and misunderstanding in responding to the questions, MARSI was translated into a Chinese Version by the researcher.

Section 2: GEPT Test. In order to identify students' English reading comprehension, the reading section of a General English Proficiency Test (GEPT) was used in the study. The GEPT is divided into five levels according to difficulty: elementary, intermediate, high-intermediate, advanced, and superior. The GEPT elementary level was chosen in the study because it is designed for examinees who have achieved at least a junior high school level proficiency. The GEPT reading test has a total of 35 items dealing with three components of reading: vocabulary and structure, cloze texts, and reading for comprehension. Each item contains a statement that requires examinees to choose one answer that best fits its description. The reading test requires 35 minutes for participants to complete, and the maximum score is a total of 120 points.

3.3 Data Collection and Analysis

Participating students were given relevant materials including a copy of the MARSI and the GEPT reading comprehension test. Participants were guaranteed that all data

and information was collected anonymously and would not be accessed by anyone other than the researcher.

The data was gathered from the survey with a five-point Likert scale, and from the GEPT reading scores. Before data analysis, the researcher checked and edited the data from returned questionnaires. The Statistical Package for the Social Sciences (SPSS), Version 16.0, was used for data analysis in this study.

4 Research Findings

The following information, including descriptive statistics and analysis summary, was described by the research questions of the study.

4.1 Students' Reading Comprehension

On average reading scores, as shown in Table 1, revealed that Occupational Therapy groups scored the highest (M = 67.87, SD = 18.926) on English reading comprehension than the other four groups (N: M = 39.61, SD = 18.236; PT: M = 51.28, SD = 24.878; D: M = 62.18, SD = 14.968; E: M = 67.75, SD = 20.000).

The results showed that the occupational therapy groups had better English reading scores, followed by applied English groups, dental laboratory technology and physical therapy, whereas nursing groups reported the lowest reading scores.

Table 1. Descriptive Statistic of English Reading Scores with Different Majors

Major	N	M	SD
Overall	454	55.72	23.156
Nursing (N)	143	39.61	18.236
Physical Therapy (PT)	70	51.28	24.878
Occupational Therapy (OT)	34	67.87	18.926
Dental Laboratory Technology (D)	52	62.18	14.968
Applied English (E)	155	67.75	20.000

4.2 Students' Metacognitive Awareness of Reading Strategy Use

As shown in Table 2, on average students' overall and the three types of reading strategies reported medium use of reading strategies when reading materials (M = 2.5~3.4). In term of individual strategy use, students used problem-solving strategies the most (M = 3.36, SD = .807), followed by global strategies (M = 3.30, SD = .719), whereas support strategies are used the least (M = 3.18, SD = .782). Moreover, the overall frequency of strategy use in the study was in the medium use (M = 3.28, SD = .732).

Table 2. Descriptive Statistics of the Use of Reading Strategies

Type of strategy	M	SD
Overall	3.28	.732
Global	3.30	.719
Problem-solving	3.36	.807
Support	3.18	.782

Note. $M \leqq 2.4$ indicates low use of strategies while reading; $2.5 < M < 3.4$ indicates medium use of strategies while reading; $M \geqq 3.5$ indicates high use of strategies while reading.

4.3 The Relationship between Metacognitive Awareness of Reading Strategies and Reading Comprehension

A standard regression analysis (see Table 3) was used between the dependent and independent variables. In terms of individual variable, major ($\beta = .334$, $p < .001$) and overall strategy use ($\beta = .420$, $p < .001$) both held significant relationships with students' reading scores. According to the results in Table 3 and 5, major difference is a factor that affects students on their English reading achievement. For the use of overall reading strategies, students who used more reading strategies scored better in English reading than those who used fewer strategies.

Table 3. Regression Analysis Summary for Major, Grade, Gender, Overall Reading Strategy Predicting English Reading Scores

Variables	*B*	*SEB*	*β*	*t*
Major	2.235	.294	.334	7.608***
Grade	1.004	1.639	.026	.613
Gender	-1.775	2.109	-.032	-.841
Overall strategy	13.301	1.231	.420	10.801***

Note. $R^2 = .372$, Adjust $R^2 = .366$. $F_{(4,449)} = 66.425$, $p < .001$.
*$p < .05$. ** $p < .01$. *** $p < .001$.

Another regression analysis (see Table 4) was employed to determine the relationships of three types of reading strategy use and English reading achievement. In terms of individual variable, major ($\beta = .333$, $p < .001$) and problem-solving strategy use ($\beta = .263$, $p < .001$) both held significant relationships with students' reading scores. According to the results in Table 3 and 6, major difference still presented effects on students' English reading achievement. Students in OT, D, and E groups, who used more strategies while reading, achieved higher reading scores than N and PT groups. For the three types of reading strategy, students who used more problem-solving reading strategies scored better in English reading than those who used fewer problem-solving strategies.

Table 4. Regression Analysis Summary for Major, Grade, Gender, Three types of Reading Strategy Predicting English Reading Scores

Variables	*B*	*SEB*	*β*	*t*
Major	2.229	.294	.333	7.581***
Grade	1.163	1.646	.030	.707
Gender	-2.037	2.123	-.037	-.960
Reading strategy				
Globe	4.064	2.861	.126	1.420
Problem	7.535	2.459	.263	3.064**
Support	1.531	2.272	.052	.674

Note. $R^2 = .375$, Adjust $R^2 = .367$. $F_{(6,447)} = 44.720$, $p < .001$.
*$p < .05$. ** $p < .01$. *** $p < .001$.

5 Conclusion and Implication

5.1 Discussion

Students' Metacognitive Awareness of Reading Strategy Use. The overall frequency of strategy use in the study was in the medium use. In terms of reading strategy use, the strategy that Taiwanese EFL students used the most was problem-solving strategy, and this has been supported by studies (Alsheikh, 2002; Chen, 2007; Chen, 2010; Hu, 2011). Moreover, other results have also reported that support strategy was found to be the least used (Al-Nujaidi, 2003; Chen, 2007; Chen, 2010; Wu, 2005). One possible explanation for problem-solving strategies being the most used strategy could be that instruction of English reading in the first three years of junior colleges emphasizes seeking for the correct answers, and for this purpose, using additional resources (such as library and online information access) or cooperating different realm of information are not required at this stage.

Metacognitive Awareness of Reading Strategy Use Relationship to English reading Comprehension. For the use of overall reading strategies, students who used more reading strategies scored better in English reading than those who used fewer strategies. According various research findings, it was reported that the training of metacognitive reading strategy has positive effect on developing students' reading performance (Tseng, 2009; Wu, 2012). That is, reading comprehension could be promoted through metacognitive reading strategy training which helps increase the use of strategies on reading.

In terms of individual reading strategy use, students who used more problem-solving reading strategies scored better in English reading than those who used fewer problem-solving strategies. The finding can be explained by a fact that English, instead of reading for fun, is an academic subject which is used to examine students' English performance. Thus, students are taught to seek for the right answers as soon as possible. In addition, intense class schedule could not offer student enough time and opportunity to search supportive information related to the reading materials.

5.2 Implication

Effective Teaching Strategies for Reading Comprehension. In the study, students were reported to possess medium use of strategies on reading. That is, students adapted strategies to help comprehend written text, either intentionally or spontaneously. It comes to an agreement that better readers are often strategic and skillful (Celce-Murcia, 2001; Tompkins, 2005). Besides, since the 1970s, a number of models and strategies of reading comprehension have been developed. Research for the National Reading Panel has identified five effective reading comprehension strategies which are "summarization, self-questioning, story structure instruction, graphic and semantic organizer, and comprehension monitoring" (Taylor, et al., 2006, p.305).

To this point, Brown and Palincsar (1989) provided four reading strategies, called reciprocal teaching (RT), that should be taught to students; summarizing, predicting, clarifying, and asking questions. According to the research findings, reciprocal teaching has been reported a significance on promoting metacognition (Huang, 1996; Yang, 2002) and reading comprehension (Frances & Eckart, 1992; Hsieh, 2010; Lin, 2012; Tsai, 2010; Ya, 2010). Since English reading is often taught as an academic subject in most Taiwanese classes and finding the answer is always the only mission to read. That is, it left no need to students to probe the information behind the written text and then to connect it to their prior knowledge. Therefore, it is necessary to offer students the training and practice about using the four types of RT. Asking questions, for instance, is one of the most common modes to engaging responsively. Different levels of questions lead to different levels of cognitive engagement with text. It is found that teachers who use more high-level questions significantly improve students' reading comprehension (Arends, 1994; Rothenberg & Fisher, 2007; Taylor, Pearson, Clark, & Walpole, 2000; Taylor, et al., 2006; Wilen, 1991).

Acknowledgements. I would like to show my gratitude to Shu Zen College of Medicine and Management and faculty of Department of Applied English. Without their support, this work could not have been done. This work was supported in part by Shu Zen College of Medicine and Management under the Grants SZE10106011.

References

1. Arends, R.: Learning to teach. McGrew-Hill, New York, NY (1994)
2. Al-Nujaidi, A.H.: The relationship between vocabulary size, reading strategies, and reading comprehension of EFL learners in Saudi Arabia. Unpublished doctoral dissertation. Oklahoma State University, Stillwater (2003)
3. Alsheikh, N.O.: An examination of the metacognitive reading strategies used by native speakers of Arabic when reading academic texts in Arabic and English. Oklahoma State University, Stillwater (2002)
4. Baker, L., Brown, A.L.: Metacognitive Skills and Reading. In: Pearson, P.D., Barr, R., Kamil, M.L., Mosenthal, P. (eds.) Handbook of Reading Research, pp. 353–394. Longman, New York (1984)

5. Brown, A., Palincsar, A.: Guided, cooperative learning and individual knowledge acquisition. In: Resnick, L.B. (ed.) Knowledge, Learning and Instruction: Essays in Honor of Robert Glaser, pp. 393–451. Lawrence Erlbaum, Hillsdale (1989)
6. Celce-Murcia, M.: Teaching English as a second or foreign language, 3rd edn. Heinle & Heinle, Boston (2001)
7. Chen, C.H.: Metacognitive Reading Strategies Used by College students and Their FL Reading Attitudes. Unpublished master thesis. Southern Taiwan University of Science and Technology, Tainan, Taiwan (2010)
8. Chen, L.C.: A study of the relationship between EFL reading anxiety and reading strategy use. Unpublished master thesis. National Taiwan University of Science and technology, Taipei, Taiwan (2007)
9. Cook, V.: Second language learning and language teaching, 3rd edn. Arnold, London (2001)
10. Farley, M.J., Elmore, P.B.: The relationship of reading comprehension to critical thinking skills, cognitive ability, and vocabulary for a sample of underachieving college freshmen. Educational and Psychological Measurement 52, 921–931 (1992)
11. Flavell, J.H.: Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. American Psychologist 34, 906–911 (1979)
12. Frances, S.M., Eckart, J.A.: The Effects of Reciprocal Teaching on Comprehension. Unpublished research project. Oakland University, Auburn Hills, MI (1992)
13. Houtveen, A., Van de Grify, W.: Instruction and Instruction Time on Reading Comprehension. School Effectiveness and School Improvement 18(2), 173–190 (2006)
14. Hsieh, Y.S.: The Effects of Reciprocal Teaching on Reading Comprehension learning efficiency in the third grade Children of the New Inhabitants. Unpublished master thesis, National Chiayi University, Chiayi, Taiwan (2010)
15. Hu, H.E.: An Investigation of Taiwanese Vocational High School Students' Use of Metacognitive Reading Strategies. Unpublished master thesis. National Taiwan Normal University, Taipei, Taiwan (2011)
16. Huang, Q.Y.: Effects of Reciprocal Teaching on Reading Comprehension Ability, Metacognitive Ability and Reading Attitude of Elementary School Sixth Grade Students. Unpublished master thesis. National Chiayi University, Chiayi, Taiwan (1996)
17. Lin, Y.C.: An Action Research of Integrating Reciprocal Teaching into Class Reading Group on Promoting Reading Comprehension for Junior High School Students. Unpublished master thesis. Tamkang University, Taipei, Taiwan (2012)
18. Maria, K.: Reading comprehension instruction: Issues and strategies. York Press, Parkton (1990)
19. Mokhtari, K., Reichard, C.A.: Assessing students' metacognitive awareness of reading strategies. Journal of Educational Psychology 94(2), 249–259 (2002), doi:10.1037//0022-0663.94.2.249
20. Rothenberg, C., Fisher, D.: Teaching English language learners: A differentiated approach. Pearson education, Columbus (2007)
21. Taylor, B.M., Pearson, P.D., Clark, K., Walpole, S.: Effective schools and accomplished teachers: Lessons about primary grade reading instruction in low-income schools. Elementary School Journal 101, 121–166 (2000), doi:10.1086/499662
22. Taylor, B.M., Pearson, P.D., Garcia, G.E., Stahl, K.A., Bauer, E.B.: Improving students' reading comprehension. In: Stahl, K.A., McKenna, M.C. (eds.) Reading Research at Work: Foundations of Effective Practice, pp. 303–315. The Guilford Press, New York (2006)
23. Tompkins, G.E.: Language arts essentials. Prentice Hall, Upper Saddle River (2005)

24. Tsai, W.R.: A Study on Applying Reciprocal Teaching to Improving Adult Students" Performances of English Reading Comprehension at Junior College's Continuing Education. Unpublished doctoral dissertation, National Chung Cheng University, Chiayi, Taiwan (2010)
25. Tseng, Y.H.: The effects of metacognitive reading strategy training on English reading comprehension and attitudes of junior high school students. Unpublished doctoral dissertation, National Chengchi University, Taipei, Taiwan (2009)
26. Wilen, W.: Questioning skills for teachers: What research says to the teacher, 3rd edn. National Education Association, Washington, DC (1991)
27. Wu, Y.L.: A study of multiple intelligences and reading comprehension. Unpublished master thesis. National Taichung University of Education, Taichung, Taiwan (2012)
28. Wu, C.P.: An investigation of metacognitive reading strategies used by EFL Taiwanese college students to comprehend familiar versus unfamiliar Chinese and English Texts. Unpublished doctoral dissertation. University of Idaho, Idaho (2005)
29. Wu, H.H.: Effects of Metacognitive Reading Strategies Training on English Reading Comprehension and Strategy Use of EFL Junior High School Students in Central Taiwan. Unpublished master thesis. National Kaohsiung Normal University, Kaohsiung, Taiwan (2012)
30. Ya, L.S.: The reciprocal teaching method into class reading groups to investigate the effects of class reading groups on eighth graders. Unpublished master thesis. National Taichung University of Education, Taichung, Taiwan (2010)
31. Yang, R.C.: Effects of Reciprocal Teaching on Reading Comprehension, Metacognition, Reading Motivation of Fifth-Grade Students. Unpublished master thesis. National Pingtung University of Education, Pingtung, Taiwan (2002)

Automated Camera Selection and Control for Better Training Support

Adrian Ilie[1] and Greg Welch[2]

[1] The University of North Carolina at Chapel Hill
[2] The University of Central Florida

Abstract. Physical training ranges have been shown to be critical in helping trainees integrate previously-perfected skills. There is a growing need for streamlining the feedback participants receive after training. This need is being met by two related research efforts: approaches for automated camera selection and control, and computer vision-based approaches for automated extraction of relevant training feedback information.

We introduce a framework for augmenting the capabilities present in training ranges that aims to help in both domains. Its main component is ASCENT (Automated Selection and Control for ENhanced Training), an automated camera selection and control approach for operators that also helps provide better training feedback to trainees.

We have tested our camera control approach in simulated and laboratory settings, and are pursuing opportunities to deploy it at training ranges. In this paper we outline the elements of our framework and discuss its application for better training support.

1 Introduction

In recent years, physical training ranges have proven instrumental in providing trainees with a way to integrate skills perfected separately in an environment that is similar to the operational environment. Training feedback is provided in the form of After Action Reviews (AARs), which currently require a large number of highly-experienced instructors to accompany different segments of the unit throughout their training run.

Some training ranges have been equipped with large networks of hundreds of cameras, which can capture training exercises as they take place. Many cameras have pan-tilt-zoom (PTZ) capabilities, and are manually controlled by operators during the exercises. Other cameras are static, but operators still need to manually select which cameras to record, because only a limited number of recording devices are usually available. To alleviate these problems, automated approaches are being pursued to augment the operators' capabilities in controlling PTZ cameras and selecting which video streams to record.

The availability of cameras and operators has enabled instructors to provide a package containing multiple hours-long video segments manually selected by the operators. However, in order to pinpoint problem areas, the videos in the

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 50–59, 2013.

package need to be reviewed in their entirety. Computer vision algorithms can be employed to analyze the captured images and automatically extract information relevant for training feedback.

The framework introduced in this paper supports these efforts through ASCENT, an automated camera selection and control approach designed to support camera operators, while also helping provide better feedback to trainees by taking into account the requirements of the computer vision algorithms that process the captured images.

ASCENT consists of a *stochastic performance metric* and a *constrained optimization method*. The performance metric quantifies the uncertainty in the state of the targets. It can account for occlusions, accommodate requirements specific to the algorithms used to process the images, and incorporate other factors that can affect their results. The optimization method explores the space of camera configurations over time under constraints associated with the cameras, the predicted target trajectories, and the image processing algorithms. To achieve real-time performance, it combines a global assignment of cameras to targets that divides the problem into subproblems with a local optimization inside each subproblem. The global assignment uses a proximity-based heuristic to group targets and a greedy heuristic based on performance metric evaluations to assign cameras to each target group. It can also perform camera selection when needed. The local optimization is performed at the level of each group. It predicts the trajectories of all targets in the group and plans dynamic camera configurations over time to ensure optimal coverage up to a time horizon. While only some of the available cameras may be selected for recording, all captured images are available for algorithms that run in real-time, some of which can even provide feedback to ASCENT.

We have applied ASCENT to simulated and laboratory settings, and are pursuing opportunities to deploy it at training ranges that have already been outfitted with large camera networks. Our framework is well-positioned to help augment training capabilities. First, it augments camera operators' capabilities, allowing them to more effectively manage large camera networks. ASCENT automates camera selection and control decisions, allowing operators to either direct it to cover important events, or directly manage a smaller number of cameras. Additionally, ASCENT can be customized to produce images best-suited for the computer vision approaches that analyze and help extract relevant training feedback data. This has the potential to shorten AAR video packages down to automatically-selected segments that can be reviewed much faster.

The rest of the paper is organized as follows. In Section 2 we present some relevant research: a few performance metrics and camera control methods, as well as a few computer vision approaches that can be used to augment training. Section 3 presents our approach to camera selection and control: our performance metric and our camera selection and control method, as well as some experimental results. Section 4 briefly describes our framework and its potential contributions to better training support. We discuss some future work and conclude the paper in Section 5.

2 Previous Work

2.1 Performance Metrics

Many researchers have attempted to express the intricacies of factors such as placement, resolution, field of view, focus, etc. into metrics that could measure and predict camera performance. Below we list the performance metrics research closest to our work. The interested reader can find a comprehensive list of camera performance metrics in Chapter 2 of [10].

Allen [1] introduces steady-state uncertainty as a performance metric for optimizing the design of multi-sensor systems. In previous work [9] we illustrate the integration of several performance factors into this metric and envision applying it to 3D reconstruction using active cameras.

Denzler et al. [3] derive a performance metric based on conditional entropy to select the camera parameters that result in sensor data containing the most information for the next state estimation. In [4], Denzler et al. present a performance metric for selecting the optimal focal length in 3D object tracking. The determinant of the a posteriori state covariance matrix is used to measure the uncertainty derived from the expected conditional entropy given a particular action. Visibility is taken into account by considering whether observations can be made and using the resulting probabilities as weights. The authors of Deutsch et al. [6,5] improve the process by using sequential Kalman filters to deal with a variable number of cameras and occlusions, predicting several steps into the future and speeding up the computation. The ASCENT performance metric presented in Section 3.1 is similar to the metric by Denzler et al., but it uses a norm of the error covariance instead of entropy as the metric value, and employs a different aggregation method.

2.2 Camera Selection and Control Methods

Camera selection and control methods are typically encountered in surveillance applications. Many are centralized approaches, based on the adaptation of scheduling policies, algorithms and heuristics from other domains to camera control. Others are distributed: decisions are arrived at through contributions from collaborating or competing autonomous agents. We list a few example methods below. The interested reader is referred to Chapter 2 of [10] for a comprehensive list.

Qureshi and Terzopoulos [19] propose a virtual testbed for surveillance algorithms and use it to demonstrate two adapted scheduling policies: first come, first serve (FCFS) and earliest deadline first (EDF). In [18], they apply the same paradigm to a distributed surveillance system, in which cameras can organize into groups to accomplish tasks using local processing and inter-camera communication with neighbors in wireless range.

Naish et al. [17] propose applying principles from dispatching service vehicles to the problem of optimal sensing. They present a dynamic dispatching methodology that selects and maneuvers subsets of available sensors for optimal data

acquisition in real-time. The goal is to select the optimal sensor subset for data fusion by maneuvering some sensors in response to target motion while keeping other sensors available for future demands.

Lim et al. [13] propose solving the camera scheduling problem using dynamic programming and greedy heuristics. The goal of their approach is to capture images that satisfy task-specific requirements such as: visibility, movement direction, camera capabilities, and task-specific minimum resolution and duration.

Krahnstoever et al. [11] present a system for controlling 4 PTZ cameras to accomplish a biometric task. Scheduling is accomplished by computing camera plans: lists of targets to cover at each time step. Plans are evaluated using a probabilistic performance objective function to optimize the task success probability.

Broaddus et al. [2] present *ACTvision*, a system consisting of a network of PTZ cameras and GPS sensors covering a single connected area that aims to maintain visibility of designated targets. Cameras are tasked to follow specific targets based on a cost calculation that optimizes the task-camera assignment and performs hand-offs from camera to camera. The authors develop optimization strategies to either use the minimum number of cameras needed, or encourage multiple views of a target for 3D reconstruction.

Sommerlande and Reid [21] present a probabilistic approach to control multiple active cameras observing a scene. Similar to the approach in ASCENT, they cast control as an optimization problem, but their goal is to maximize the expected mutual information gain as a measure for the utility of each parameter setting and each goal. The approach allows balancing conflicting goals such as target detection and obtaining high resolution images of each target.

Matsuyama and Ukita [15] describe a distributed system for real-time multi-target tracking. The system is organized in three layers (inter-agency, agency and agent), with agents that dynamically interchange information with each other.

2.3 Computer Vision Approaches to Augment Training

There are many computer vision approaches that can process images, ranging from posture recognition from single images [22] to full 3D reconstruction from multiple images: multi-view dynamic scene modeling [7], space carving [12], 3D video [14] and image-based visual hulls [16]. However, most of these approaches have yet to be applied to large environments such as training ranges. Moreover, there are few approaches that can analyze the results of computer vision algorithms and extract relevant information that can help augment training. Sadagic et. al. [20] describe a concerted research effort in this direction. ASCENT provides ways to take into account the requirements of these approaches in order to capture images that are likely to produce the best possible result.

3 Automated Camera Selection and Control

We approach camera selection and control as an *optimization problem* over the space of possible *camera configurations* (combinations of camera settings) and

over time, under constraints derived from knowledge about the cameras, the predicted target trajectories and the computer vision algorithms the captured images are intended for. The objective function is a performance metric that evaluates dynamic, evolving camera configurations over time. In this section, we briefly describe the two components of ASCENT: its camera performance metric and its camera selection and control method. The interested reader is referred to [8] and Chapters 5 and 6 of [10] for a detailed presentation.

3.1 Camera Performance Metric

We define the performance of a camera configuration as its ability to resolve 3D features in the working volume, and measure it using the uncertainty in the state estimation process. We use state-space models [8] to describe target dynamics and measurement systems. Formally, at time step t, the system state is described by a *state vector* $\bar{x}_t \in \mathbb{R}^n$ which may include elements for position, orientation, velocity, etc. Given a point in the state space, a mathematical *motion model* can be used to predict how the target will move over a given time interval. Similarly, a *measurement model* can be used to predict what will be measured by each sensor. We measure the uncertainty in the state $\bar{x}_t$ using the a posteriori error covariance P_t^+, which we compute by applying the Kalman Filter equations to elements of the state-space models.

Our performance metric evaluates *plans*: temporal sequences of camera configurations up to a *planning horizon*. We compute the performance metric for each candidate plan by repeatedly stepping forward in time up to the planning horizon, while applying the Kalman Filter equations and changing relevant state-space model parameters at each time step. We use the motion models to predict target trajectories and generate predicted measurements, and we update the measurement models with the camera parameters corresponding to the configurations planned for each time step. We aggregate over space and time using weighted sums, with weights quantifying the relative importance of elements at various levels, such as points in a target surrogate model, targets, or time instants. Equation 1 illustrates the general formula for the metric computation using weighted sums.

$$\mathcal{M} = \sum_{r=1}^{N_t} u_r \left(\sum_{t=1}^{H} v_t \left(\sum_{p=1}^{N_r} w_p \left(\sqrt{Max\left(Diag_{pos}\left(P_{t,p}^+\right)\right)} \right) \right) \right) \tag{1}$$

N_t is the number of targets, N_r is the number of points in the surrogate model of target r, H is the planning horizon. u_r, v_t and w_p are relative weights for each target r, time step t, and model point p, respectively. $P_{t,p}^+$ is the a posteriori covariance for model point p at time t. To convert the error covariance into a single number, we use the square root of the maximum value on the diagonal of the portion of the error covariance matrix $P_{t,p}^+$ corresponding to the position part of the state.

3.2 Camera Selection and Control Method

We define optimization in active camera selection and control as the exploration of the space of possible solutions in search for the best solution as evaluated by the performance metric. Our optimization process first predicts the target trajectories, then uses them to construct and evaluate a number of candidate plans for each camera. A plan consists of a number of *planning steps*. A step consists of a *transition* (during which cameras are not being recorded, and PTZ cameras change their settings) and a *dwell* (during which cameras capture, with constant settings, and are being recorded). Candidate plans differ in the number and duration of planning steps up to the planning horizon.

To ensure real-time performance, we decompose the optimization problem into subproblems and solve each subproblem independently. Our method consists of two components: centralized *global assignment* and distributed *local planning*.

The global assignment component accomplishes two tasks: grouping targets into *agencies* and assigning cameras to each agency. We create agencies by clustering together targets that are close to each other and predicted to be heading in the same direction. We use predicted target trajectories to cluster the targets into a minimum number of non-overlapping agencies of a given maximum diameter. We use a *minimal change* clustering heuristic that tries to preserve agency membership over time. We then use a greedy heuristic to assign cameras to each agency, based on their potential contribution to it. The heuristic iteratively tries assigning all available cameras to nearby agencies, searching for the camera-agency assignment that best improves the performance metric value for the agency. Improvement is measured using the ratio between the metric values before and after making the assignment. The resulting plans are compared with plans obtained by prolonging the current plans up to the planning horizon whenever possible, and the greedy assignments are only applied if they perform better. We use the same process both to control PTZ cameras in real-time and to select which cameras to record when there are fewer recording devices than cameras. In the case of selection, we simply stop after the maximum allowable number of cameras have been assigned. The plans corresponding to each camera-agency assignment are generated assuming the worst-case scenario: the camera is repeatedly set to transition, then capture for as long as possible, with PTZ cameras zoomed out to a field of view as wide as possible. Predicted static and dynamic occlusions are taken into account, and transitions are planned during occlusions whenever possible, in order to minimize the time intervals when cameras are not capturing.

Local planning at the level of each agency is concerned with the locally-optimal capture of the targets in the agency. All cameras assigned to each agency capture all member targets, and no further camera-target assignment decisions are made at this level. The planning decisions made at this level are on when and for how long each camera should dwell (capture), and when each PTZ camera should transition to a new configuration. All possible combinations of candidate plans for all cameras are explored exhaustively using backtracking. To achieve on-line, real-time control, the set of candidate plans is heuristically generated

and sorted so that the most promising plans are evaluated first. We use prior experimental observations to derive criteria for judging a plan's potential. While not a guarantee that the best plan would be chosen on time, we have found this heuristic to closely approximate an exhaustive search.

3.3 Experimental Results

We have applied ASCENT to automated on-line control of cameras in simulated and laboratory settings, capturing training exercises that involved patrolling, cordoning and searching a civilian, and crossing a danger zone. Experiments showed the emergence of desired camera behaviors, including: fast coverage of new targets, continuous target coverage via staggered settings adjustments, continuous coverage of divergent target groups, automatic hand-offs, and continuous preemptive coverage of fast-moving targets. The performance metric and control method were tuned to produce images best suited for a volumetric reconstruction method such as [7]. The interested reader is referred to Chapter 7 of [10] and [8] for more details.

The simulated setting involved capturing 6 targets (4 Marines and 2 civilians) moving around 2 occluders, using 6 cameras. Figure 1 (Left) shows an overview of the setup as modeled in the simulator. Camera locations are shown in blue, occluders are shown in red. The laboratory setting involved capturing 7 targets (4 Marines and 3 civilians) moving around the entrance to an alley between 2 buildings, using 8 cameras hanging from the ceiling. Figure 1 (Right) shows an image captured by an overview camera during the exercise. Building walls were simulated using cloth attached to waist-high posts.

Fig. 1. (Left) Simulated setting. (Right) Laboratory setting.

4 Training Support Framework

We envision ASCENT as part of a training support framework that defines how automatic control of cameras can augment the capabilities at training ranges.

First, by automating camera selection and control decisions, ASCENT augments the operators' capabilities. A well-configured automated system can make decisions that an operator may find counter-intuitive, but are justified when the

captured images are destined for automated analysis, as opposed to manual review. We envision the following scenarios for how ASCENT can be applied to augment human decisions in camera selection and control:

1. *Automated:* ASCENT controls all active cameras and selects a number of cameras for recording.
2. *Directed:* ASCENT allows operators to intervene on-the-fly, and designate important events, areas and persons for capture with higher priority. Camera selection and control are still done automatically, but operator interventions are incorporated as constraints in the optimization method.
3. *Assisted:* ASCENT allows operators to dynamically choose a set of cameras that they want to record, control directly or assign to particular targets. It then assists the operators by automatically selecting which of the remaining cameras to record, and controlling the remaining active cameras. It also suggests the best camera-target assignments and camera settings for the cameras chosen by the operators, but lets the operators decide whether to apply them or not.

Second, ASCENT augments the training capabilities at training ranges by helping provide images best suited for automated computer vision analysis, which has the potential to shorten AARs video packages down to segments relevant for improving the trainees' performance. To that end, both components of ASCENT are highly customizable. The performance metric can be adapted to include performance factors relevant to the application, such as varying weights for different members of a team over time; or factors relevant to the computer vision algorithm used, such as preferred incidence angles for 3D reconstruction or 2D posture recognition. The selection and control method can incorporate domain knowledge such as the training range topology and the locations of important training events in relation to camera placement, as well as their timing during a training exercise. The interested reader can find a discussion of many of the customizations possible in ASCENT in [8] and Chapters 5 and 6 of [10].

5 Conclusions and Future Work

We introduced a framework for augmenting capabilities at training ranges. Its main component is ASCENT, an optimization-based on-line camera selection and control approach consisting of a performance metric and a selection and control method. For the optimization objective function, we employ a versatile performance metric that can incorporate both camera performance factors and application requirements. To reduce the size of the search space and arrive at an implementation that runs in real-time, our camera control method breaks down the optimization problem into subproblems. We first use a proximity-based minimal change heuristic to decompose the problem into subproblems and a greedy heuristic to select cameras and assign them to subproblems. We then solve each subproblem independently, generating and evaluating candidate plans as time allows. We applied ASCENT to simulated and laboratory settings, demonstrating

useful camera behaviors. We briefly discussed how ASCENT can help augment the capabilities at training ranges: it can automate selection and control decisions, and can be easily adapted to include requirements for automated analysis using computer vision approaches.

We are looking forward to applying ASCENT in training ranges that have the camera infrastructure already in place, and gather feedback from camera operators, instructors and trainees. We plan to address the challenges of scaling an approach that has only been tested in simulated and laboratory settings with a small number of cameras to training ranges with hundreds of cameras. We are also looking forward to incorporating the requirements of emerging approaches that go beyond the results of today's computer vision algorithms and extract relevant information such as the video segments best suited for AARs. While in its current version ASCENT can capture images best suited for computer vision, human reviewers may have different requirements for AAR. We plan to leverage the experience of human operators in selecting footage appropriate for AARs in further customizing ASCENT to incorporate these requirements. Similarly, the experience of instructors currently following monitoring exercises on the ground will be invaluable.

Acknowledgments. We acknowledge our sponsors and collaborators in the "Behavior Analysis and Synthesis for Intelligent Training (BASE-IT)" project: ONR grant N00014-08-C-0349, Roy Stripling, Ph.D., Program Manager, led by Amela Sadagic (PI) at the Naval Post-graduate School, Greg Welch (PI) at UNC, and Rakesh Kumar (PI) and Hui Cheng (Co-PI) at Sarnoff.

References

1. Allen, B.D.: Hardware Design Optimization for Human Motion Tracking Systems. Ph.D. thesis, University of North Carolina at Chapel Hill (December 2007)
2. Broaddus, C., Germano, T., Vandervalk, N., Divakaran, A., Wu, S., Sawhney, H.: Act-vision: active collaborative tracking for multiple ptz cameras. In: Proceedings of SPIE: Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications, Orlando, FL, USA, vol. 7345 (April 2009)
3. Denzler, J., Zobel, M., Niemann, H.: On optimal camera parameter selection in kalman filter based object tracking. In: 24th DAGM Symposium on Pattern Recognition, pp. 17–25 (2002)
4. Denzler, J., Zobel, M., Niemann, H.: Information theoretic focal length selection for real-time active 3-d object tracking. In: International Conference on Computer Vision, vol. 1, pp. 400–407 (October 2003)
5. Deutsch, B., Niemann, H., Denzler, J.: Multi-step active object tracking with entropy based optimal actions using the sequential kalman filter. In: IEEE International Conference on Image Processing, vol. 3, pp. 105–108 (2005)
6. Deutsch, B., Zobel, M., Denzler, J., Niemann, H.: Multi-step entropy based sensor control for visual object tracking. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 359–366. Springer, Heidelberg (2004)

7. Guan, L.: Multi-view Dynamic Scene Modeling. Ph.D. thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA (April 2010)
8. Ilie, A., Welch, G.: On-line control of active camera networks for computer vision tasks. ACM Transactions on Sensor Networks (to appear, 2014)
9. Ilie, A., Welch, G., Macenko, M.: A stochastic quality metric for optimal control of active camera network configurations for 3D computer vision tasks. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications, Marseille, France (October 2008)
10. Ilie, D.A.: On-Line Control of Active Camera Networks. Ph.D. thesis, University of North Carolina at Chapel Hill (2010)
11. Krahnstoever, N., Yu, T., Lim, S.N., Patwardhan, K., Tu, P.: Collaborative real-time control of active cameras in large scale surveillance systems. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications, Marseille, France (October 2008)
12. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International Journal of Computer Vision 38(3), 199–218 (2000)
13. Lim, S.-N., Davis, L., Mittal, A.: Task scheduling in large camera networks. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part I. LNCS, vol. 4843, pp. 397–407. Springer, Heidelberg (2007)
14. Matsuyama, T., Wu, X., Takai, T., Nobuhara, S.: Real-time 3d shape reconstruction, dynamic 3d mesh deformation, and high fidelity visualization for 3d video. In: Computer Vision and Image Understanding, vol. 96, pp. 393–434. lsevier Science Inc., New York (2004)
15. Matsuyama, T., Ukita, N.: Real-time multitarget tracking by a cooperative distributed vision system. Proceedings of the IEEE 90, 1137–1150 (2002)
16. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: ACM Siggraph, pp. 369–374 (2000)
17. Naish, M.D., Croft, E.A., Benhabib, B.: Coordinated dispatching of proximity sensors for the surveillance of manoeuvring targets. Robotics and Computer-Integrated Manufacturing 19(3), 283–299 (2003)
18. Qureshi, F., Terzopoulos, D.: Surveillance in virtual reality: System design and multi-camera control. In: Proceedings of Computer Vision and Pattern Recognition, pp. 1–8 (June 2007)
19. Qureshi, F.Z., Terzopoulos, D.: Towards intelligent camera networks: a virtual vision approach. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, pp. 177–184 (October 2005)
20. Sadagic, A., Welch, G., Basu, C., Darken, C., Kumar, R., Fuchs, H., Cheng, H., Frahm, J.M., Kolsch, M., Rowe, N., Towles, H., Wachs, J., Lastra, A.: New generation of instrumented ranges: Enabling automated performance analysis. In: 2009 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC-2009), Orlando, FL (2009)
21. Sommerlade, E., Reid, I.: Probabilistic surveillance with multiple active cameras. In: IEEE International Conference on Robotics and Automation (May 2010)
22. Wachs, J., Goshorn, D., Kolsch, M.: Recognizing human postures and poses in monocular still images. In: Intl. Conf. on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, NV, pp. 665–671 (2009)

A Hierarchical Behavior Analysis Approach for Automated Trainee Performance Evaluation in Training Ranges

Saad Khan, Hui Cheng, and Rakesh Kumar

SRI International, Princeton, USA
{saad.khan,hui.cheng,rakesh.kumar}@sri.com

Abstract. In this paper we present a closed loop mixed reality training system that provides automatic assessment of trainee performance during kinetic military exercises. At the core of our system is a hierarchical behavior analysis approach that integrates a number of data sensor modalities including Audio/Video, RFID and IMUs to automatically capture trainee actions in a comprehensive manner. Our behavior analysis and performance evaluation framework uses a finite state machine (FSM) model in which trainee behaviors are the states of the training scenario and the transitions of states are caused by stimuli that we refer to as trigger events. The goal of behavior analysis is to estimate the states of the trainees with respect to the training scenario and quantify trainee performance. To robustly detect each state, we build classifiers for each behavioral state and trigger event. At a given time, based on the state estimation, a set of related classifiers are activated for detecting trigger events and states that can be transitioned to and from the current states. The overall structure of the FSM and trigger events is determined by a Training Ontology that is specific to the training scenario.

1 Introduction

Infantry training, from basic training at home stations to joint exercises prior to deployment, can become more effective through automated behavior analysis and performance evaluations. In this paper, we present an automated behavior analysis and performance evaluation computational framework for a wide range of training objectives.

We model trainee behavior (individually and in teams) as states, and the causes of state transitions as trigger-events. Each state has a set of performance metrics. The overall goals of the training exercise are captured as hierarchical Finite State Machines (FSM) with associated performance metrics. Our behavior analysis module uses sensor data as observations to estimate the states that the trainees are in. The performance evaluation module computes the performance metrics given the estimated states of the trainees. Trigger events that result in transition from one state to another are detected using a Histograms of Oriented Occurrence (HO2) algorithm for

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 60–69, 2013.

individual and group activity recognition that captures the interactions of multiple players in one feature vector.

The system uses a suite of multi-modal sensors to capture training exercise (see figure 1). Each trainee's location, weapon and head orientations are computed using a combination of GPS or RFID, inertia navigation sensors (INS) data and video analysis. Gunshots are captured through trigger sensors and laser shot detection system. Videos and detected events are overlaid on a 3D-model of the training site for enhanced AAR and situational awareness experiences. Additionally, our AAR allows searching and browsing of training events and the computation of statistics. Our system estimates behaviors and corresponding performance metrics in real-time, and ingests both into a database. Experimental results from out prototype training system have shown improved training efficiency and effectiveness as a result of the system.

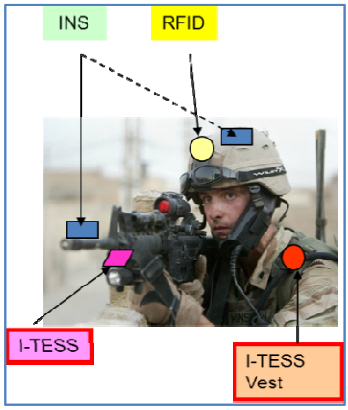


Fig. 1. Multi-modal sensor suite used to instrument trainees

2 Automated Behavior Analysis and Event Detection

Our training domain (military exercises) circumscribes the space of observable behaviors to a controllable list. Taking advantage of this fact we are able to develop a behavior analysis framework that uses a finite state machine (FSM) model where participants' behavior are the *states* and the transitions of states are caused by stimuli that we refer to as *trigger events*. The goal of behavior analysis is to estimate the states of the participants and the states that the participants should be in at any given time. The former are used for exercise and scenario control and the later are used for performance evaluation. To robustly detect each state, we build classifiers for not only for each state, but also for each trigger event. At a given time, based on the state estimation, a set of related classifiers are activated for detecting trigger events and states that can be transitioned to and from the current states.

We model a training exercise as a finite state machine (FSM). A FSM is a quintuple $(\Sigma, S, s_0, \delta, F)$, where:

- Σ is the input alphabet (a finite and non-empty).
- S is a finite, non-empty set of states.
- s_0 is an initial state, an element of S.
- δ is the state-transition function that returns a set of transition probabilities: $\delta : S \times \Sigma \rightarrow P(S)$.
- F is the set of final states, a subset of S.

For training,

- Σ is the set of stimuli or trigger events
- *S* is the set of possible behaviors, i.e. states of the participants.
- s_0 is an initial state.
- δ is the reaction to a stimulus. δ contains both the correct reactions to stimuli defined in a TTP and incorrect reactions that need to avoid.
- *F* is the end state of a training exercise.

For a training system, states *S* can only be perceived through sensor observations, *O*. Then, behavior analysis is to estimate states $S=\{s_0,s_1,..,s_n\}$ given sensor observation $O=\{o_0,o_1,..,o_n\}$. In our system, the sensor inputs include positions of all participants, their head, body and gun poses and shot/hit data (figure 1). However the definition of state *S* and transition trigger events depends on the Training Ontology discussed next.

2.1 Training Ontology

The training ontology captures knowledge related to a set of training objectives including TTP (Techniques, Tactics and Procedures), training scenarios and performance metrics. This is a machine understandable graphical-representation of the TTP that includes comprehensive data on scenario context, parameters for behavior recognition, and expected performance evaluation thresholds. Our training taxonomy is divided into two sub-hierarchies – a set of concepts representing states (nouns) and a set representing trigger events (verbs). Using Protégé [Noy, 2001], we assign a node to each state, along with the corresponding definition. Similarly, we assign a node to each trigger event and its definition. All states and trigger events form the taxonomy in our training ontology. For each state, we also store associated attributes including classifier and the performance metrics for the state. For each state and a given trigger event, the ontology also captures all states that it can transition to. Figure 2 illustrates an example exercise model that includes speech and gestures so that they can be assessed in the same overall framework.

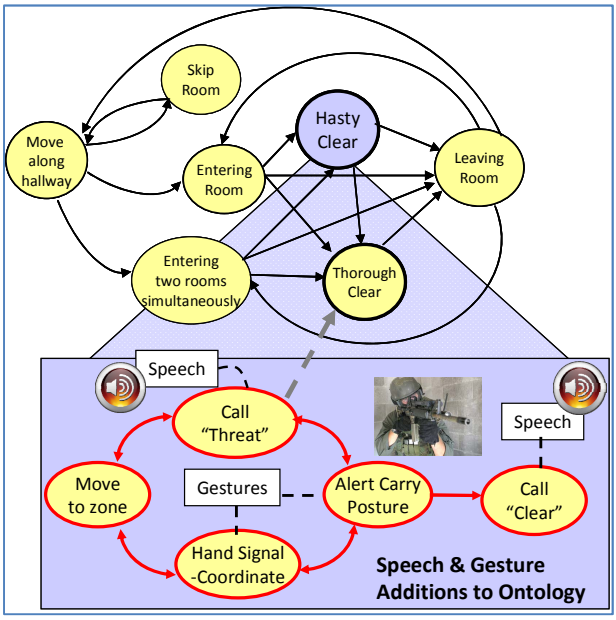


Fig. 2. Example Ontology of Exercise Room Clearing

2.2 Hierarchical Behavior Analysis

The training ontology helps us define the FSM that represents only the top layer of our hierarchical behavior analysis module. As illustrated in figure 3 at the lowest level is the Action Detection module that classifies atomic actions performed by the participants. These atomic actions span a wide array of low-level trainee behaviors like "walking", "group formation", "weapon sector scanning", "weapon fire" etc. In most cases classifiers for these atomic actions are trained on static features extracted directly from the raw sensor data. For instance to detect "group formations" the track locations of the trainees are used to match against a shape template pertaining to a "diamond" or "wedge" formation. In the middle layer, we generate Trigger Events which are mid-level abstractions of trainee behavior that result in a meaningful transition from one state in the scenario to another. These trigger events typically represent a dynamic activity that require features to be extracted over a window of time frames. Figure 3 illustrates some examples of these including "Cordon Formation", "Crossed Danger Zone" etc.

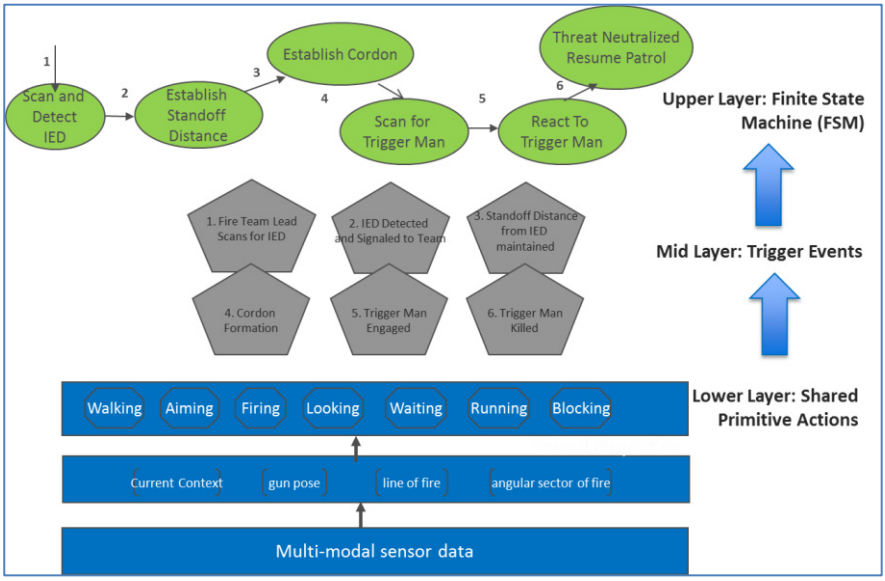


Fig. 3. Hierarchical framework for behavior analysis

Adaptive space-time aggregated features Histogram of Oriented Occurrences (HO2) are computed and trained with SVM to classify atomic actions and trigger events. In its most generalized form, space-time context is the histogram of occurrences of entity classes of interest over a partition of a spatial-temporal volume with respect to a reference entity or a reference location. Existing activity or event exploitation approaches represent these events using features that only measure pair wise relationships between entities at a time, such as relative distance and relative speed. Due to the limitations of the pair wise entity relationship descriptors, this class of

events is mainly defined and recognized using rule-based approach. HO2 captures the interactions of all entities of interests in terms of configurations over space and time through a histogramming process. Using this new space-time context representation, our activity exploitation approach captures both environmental context and spatial-temporal characteristics of the entities in a unified framework. Using HO2, we have been able to detect multi-agent events such as VIP arriving or depicturing with security details.

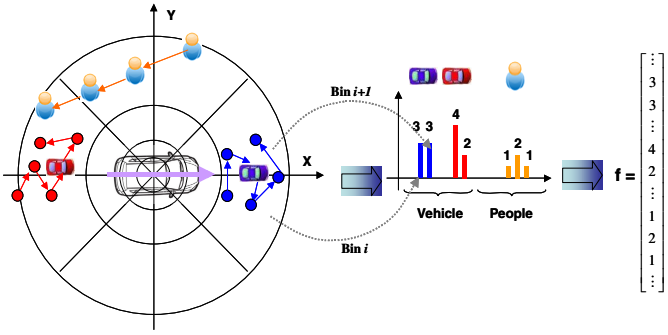


Fig. 4. HO2 computation using log-polar partition function. The reference entity is the middle vehicle in a three-vehicle convoy. The people icons represent a pedestrian crossing the street. The histograms of vehicle and people occurrences are shown in the middle. The resulting space-time context feature vector is shown on the right.

Finally, as already discussed at the third and highest level a finite state machine (FSM) is used to model the training scenario as a set of behavioral states predicated with trigger events (mid-level). The overall structure of the FSM and trigger events is determined by a Training Ontology that is specific to the TTP (techniques tactics and procedures) of the training scenario.

2.3 Trainee Performance Evaluation

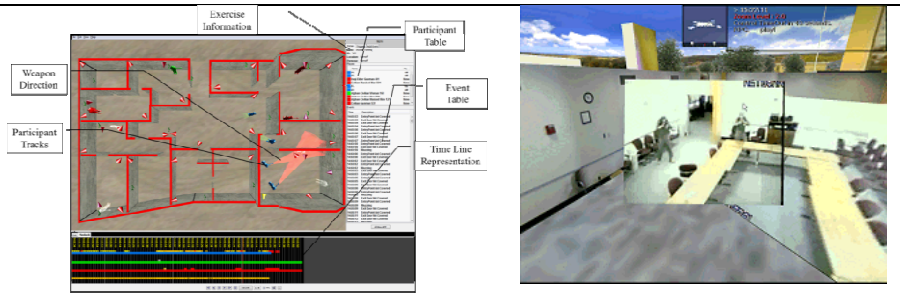
Performance metrics are computed by comparing trainee actions to canonical executions based on the TTP. Our system computes performance metrics associated with each state during a training exercise. Low-level data including location, weapon orientation etc. is used to compute these metrics. For our MOUT application training, the following performance metrics are computed:

- **360 degrees Security:** The percentage of a full 360 degrees that is either covered by a Warfighter's weapon or is blocked by a cover.
- **Blocking:** The fraction of the time that all danger spots were blocked by the warfighters, i.e. at least one warfighter points his weapon at each of the danger spots. The danger spot may be a possible sniper position or an approaching vehicle, etc. We use "Aim Margin" to determine the blocking accuracy which needs to be achieved.

- **Cover:** The fraction of time that all warfighters maintain cover. The source of cover can be natural objects such as trees, ravines, hollows, reverse slopes, etc. or man-made such as vehicles, trenches, and craters." [USMC, 2006]. The Warfighters are maintaining cover if the minimum distance of each warfighter from any of the source of cover against the threat direction is below the 'Cover Margin'. The sources of cover are computed from the 3D-model of the training environment. We use Hausdorff distance as the distance measure.
- **Flagging/Muzzling:** A warfighter points his weapon at a friendly. The Flagging score is the total number of detected flaggings.
- **Dispersion Measure:** The average nearest-neighbor distance (NND) per unit time for the warfighter team. Depending on the context, dispersion measure can be useful for signaling "bunching". For instance, "bunching" may be preferred in an urban context when a unit approaches the corner of a building; it is dangerous in open spaces in a rural context where the entire unit may be exposed.

3 Visualization and AAR

We have developed visualization tools that allow a user to view not only videos captured during an exercise, but also tracks trainees, and events in an interactive and easy-to-use manner. Two displays are provided by our system and they are used simultaneously in a synchronized fashion. They are (1) Symbolic Map Display and (2) Video Flashlight Display.



(a) Symbolic Map Display for AAR (b) Video Flashlight Display for AAR

Fig. 5. Visualization and AAR

3.1 Symbolic Map Display

In Figure 5(a), we show a snapshot of a typical exercise as it is displayed by the Symbolic Map Display. The left side of the Symbolic Map Display shows a 3D-model of the MOUT environment. The Symbolic Map Display allows users to view an exercise at any instant and track movements forward and backward in time. A user can also drag the red time line to any location and play back from there.

3.2 Video Flashlight Display

Our system uses multiple video cameras to cover a MOUT facility and captures the entire exercise. For this purpose we developed the Video Flashlight Display system [Kumar, 2003, Hsu, 2000] that can seamlessly integrate multiple video streams into a unified live display. Each of the videos is projected on the 3D-model of the environment. An example of the Flashlight Video display is show in Figure 5(b), where two video streams are projected onto the 3D-model of the scene.

4 Experiments

Our prototype training system has been tested on mock warfighter training exercises including both indoor room clearing and outdoor patrol scenarios. Two of the outdoor training exercises that we have focused on are “React to IED Threat” and “Patrolling”. Figure 5 shows an illustration of one of the simulated “React to IED Threat” exercises performed by a group of eight trainees. To perform the exercise correctly first the Marine unit needs to move towards the building to obtain cover. Once in cover one team blocks the sniper with their weapons while the other team moves across to engage and neutralize the threat.

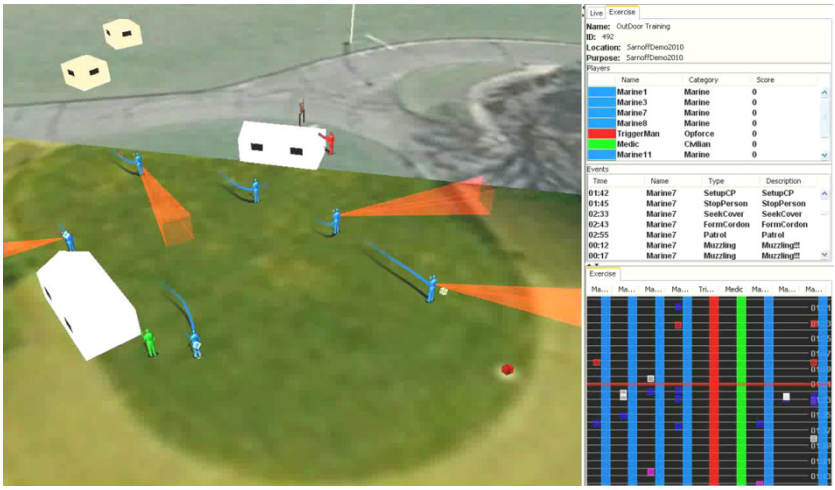


Fig. 6. Bird's eye view illustration of a "React to IED threat" exercise

In Figure 6 we show the metrics computed by our system for the “React to IED” exercise illustrated in Figure 5 (exercise 343). For “Flagging” and “Blocking” we used an “Aim Margin” parameter of 10 degrees. The “Cover Margin” parameter for “Cover Score” was set at 2 meters for this exercise. These values were selected based on empirical testing. A number of observations can be drawn from the plot in Figure 6. Only one “High Dispersion” event was detected indicating that the trainees maintained good nearest neighbor distance. On the other hand there is large number of “Faulty Blocking” and “Faulty Cover” events detected. An interesting comparison is

to view performance of two different teams on the same exercise. In figure 7 we show performance metrics comparing two teams doing the same exercise. Such comparisons are extremely useful in evaluating the impact of training and identify what metrics are more pertinent than others.

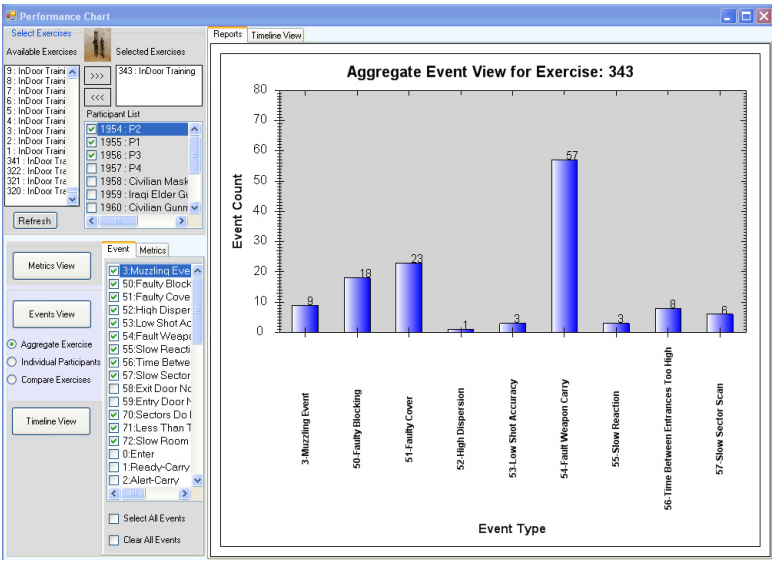


Fig. 7. Performance metrics for an exercise. Events corresponding to metrics like "Muzzling", "Cover" and others are shown.

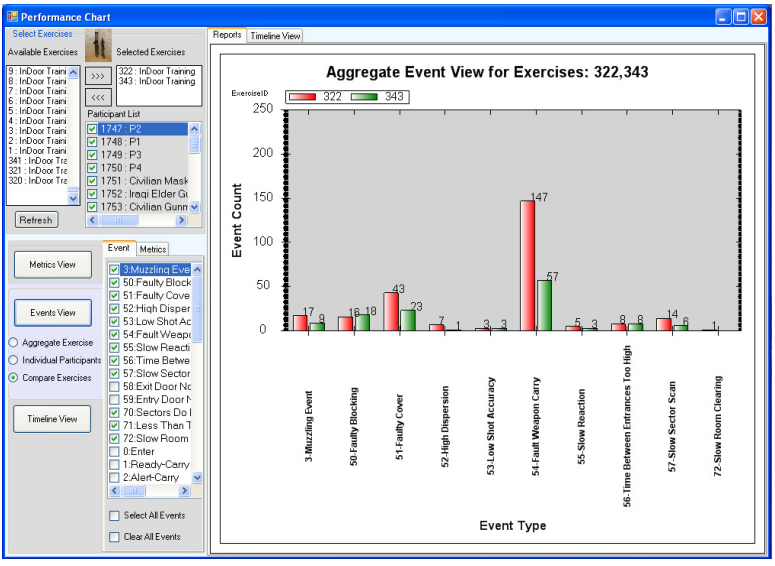


Fig. 8. Comparison between two different teams performing the same exercise

5 Conclusions

We have developed a computational framework for automated behavior analysis and performance evaluation that effectively incorporates TTP and designed training scenarios. Our approach is to use a hierarchical framework that uses a FSM at the top level to capture TTP objectives. Trigger events that transition the state machine from one state of the scenario to another are detected using classifiers on the HO2 feature. To capture trainee behavior, the prototype training system captures and computes tracks, poses and actions of the participants and automatically assesses the performance of warfighters using a training ontology. We have developed a prototype system that has been demonstrated to accurately detect participants' states, mistakes, such as muzzling, automatically. The detected events and computed performance metrics provide power tools for advanced AAR capabilities.

Acknowledgments. This work has been supported by the Office of Naval Research (ONR) program BASE-IT contract N00014-08-C-0127. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ONR, or the U.S. Government.

References

1. Cheng, H., Yang, C., Han, F., Sawhney, H.: HO2: A new feature for multi-agent event detection and recognition. In: Computer Vision Pattern Recognition Workshop, pp. 1–8 (2008)
2. Hsu, S., Samarasekera, S., Kumar, R., Sawhney, H.S.: Pose Estimation, Model Refinement, and Enhanced Visualization Using Video. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Is., SC, vol. I, pp. 488–495 (2000)
3. Jung, S., Guo, Y., Sawhney, H., Kumar, R.: Action Video Retrieval Based on Atomic Action Vocabulary. In: Proc. ACM Int'l Conf.on Multimedia Information Retrieval, Vancouver, British Columbia (2008)
4. Cheng, H., Kumar, R., Basu, C., Han, F., Khan, S., Sawhney, H., Broaddus, C., Meng, C., Sufi, A., Germano, T., Kolsch, M., Wachs, J.: An Instrumentation and Computational Framework of Automaoted Behavior Analysis and Performance Evaluation for Infantry Training. In: Proceedings of 2009 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2009), Orlando, FL (2009)
5. Cheng, H., Kumar, R., Germano, T., Meng, C.: Automatic Performance Evaluation and Lessons Learned (APELL) for MOUT Training. In: Proceedings of 2006 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2006), Orlando, FL (2006)
6. Kumar, R., Samarasekera, S., Arpa, A., Aggarwal, M., Paragano, V., Hanna, K., Sawhney, H., Sartor, M.: Monitoring Urban Sites using Video Flashlight and Analysis System. In: GOMAC Proceedings, Tampa Florida (2003)

7. Fontana, R.J.: Recent System Applications of Short-Pulse Ultra-Wideband (UWB) Technology. IEEE Transaction on Microwave Theory and Techniques 52(9), 2087–2104 (2004)
8. Noy, N.F., Sintek, M., Decker, S., Crubezy, M., Fergersen, R., Musen, M.A.: Creating Semantic Web Contents with Protégé-2000. IEEE Intelligent Systems 16(2), 60–71 (2001)
9. Melnik, S., Garcia-Molina, H., Papepcke, A.: A Mediation Infrastructure, for Digital Library Services. ACM Digital Libraries, 123–132 (2000)
10. Viola, P., Jones, M.: Robust Real-time Object Detection. In: 2nd Intl Workshop on Statistical and Comp. Theories of Vision, Vancouver (2001)
11. Wachs, J.P., Goshorn, D., Kölsch, M.: Recognizing Human Postures and Poses in Monocular Still Images. In: Intl. Conf. on Image Processing, Computer Vision, and Pattern Recognition (IPCV) (2009)
12. Torralba, S.A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE PAMI 29(5), 854–869 (2007)
13. Camouflage, Cover and Concealment, Lesson Plan. USMC, Weapons and Field Training Battalion (January 26, 2006)
14. Zhao, T., Aggarwal, M., Kumar, R., Sawhney, H.S.: Real-time Wide Area Multi-camera Stereo Tracking. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA (2005)

Augmenting Instructional Design with State-Based Assessment

Kevin Oden

Lockheed Martin Corporation, Orlando Florida, USA
Kevin.oden@lmco.com

Abstract. The Trainee Engagement Management System (TEMS) is a technology—enabled instructional design concept that leverages state-based assessment techniques to improve training processes and outcomes. Specifically, the concept is designed to support military instructors in the delivery of empirically-supported instructional prompts to foster trainee engagement within a Computer Based Training (CBT) environment. The central theme of the concept is to augment, not replace, an instructor's abilities. By reducing workload demands on an instructor, the approach enables the delivery of personalized instruction in a one (instructor) to many (trainees) context. The TEMS concept embraces a human-system philosophy and is designed to mitigate risks typically associated with the transition of advanced technologies and concepts to field settings. In this paper we discuss those challenges and describe the basic TEMS architecture.

Keywords: Instructional System Design, Augmented Cognition, Human Systems, Computer Based Training.

1 Introduction

Under conditions of persistent conflict and mounting economic pressures military instructors are required to impart mission-critical Knowledge, Skills, Abilities (KSAs) with fewer resources. Training remains as the primary mechanism for acquiring and maintaining operational readiness across all branches of the military. Many opportunities exist within instructional design field(s) to support the efficient and affordable delivery of high quality instruction. However, a co-occurrence of change across science, technology, and Operational Environments (OEs) often disrupts the successful transition of innovative and validated instructional designs to field settings. In part, a lack of coordination across these fields contributes to a growing tension as to how 'advanced' should be defined with respect to instructional technology. Researchers from each field, in earnest, work to accomplish a shared goal - improve Operator performance; however, they tend to pursue orthogonal objectives that rarely converge to produce a field-ready solution.

In this paper, we describe a technology enabled instructional design concept that embeds monitoring and management strategies to facilitate trainee engagement in

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 70–79, 2013.

Computer Based Training (CBT) settings. The primary objective is to sharpen and sustain focused attention of trainees in the service of learning. A brief review of relevant literature and contemporary work is provided to illustrate how advanced physiological technology and learning concepts can be combined to augment instructor performance and promote training effectiveness and efficiency.

2 Human-System Components

A need for field-ready instructional technologies that incorporate best offerings from the aforementioned fields has been elevated. Cross-Cutting instructional designs that leverage sufficiently mature technologies and validated learning practices are the best candidates for transition to field settings. It is clear that training system success is build upon many human factors, but it is not obvious how human should be integrated as components within the total system. An instructional design concept that integrates/balances the following instructional system components is most likely to achieve success that can be transitioned to the field in the near-term: trainee, instructor, and instructional technology.

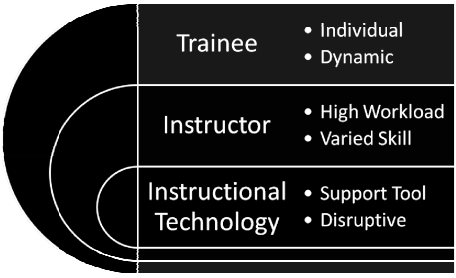


Fig. 1. Targeting Fundamental Elements of Instructional System

The components presented above (see, Figure 1) represent the essential features common across training systems. From a human-systems perspective, each component piece must be carefully considered as trade-offs are made during design. The interplay between the components is ultimately what determines total system success. Specifically, the goodness of training systems is directly proportional to a directional change in measures of trainee performance. Typically, a training system will be evaluated based on metrics of effectiveness (i.e., trainee obtains a new KSA) and efficiency (i.e., how long it took to obtain a given KSA).

Individual differences across a trainee population impacts training processes and outcomes related to effectiveness and efficiency. Each new training event is meaningfully affected by both inter-personal and intra-personal factors. For instance, a trainee's full history of learning experiences is an inter-personal factor that uniquely defines their KSA baseline. Equivalent amounts time spent on the same task may increase similarity between two trainees, but individual differences will persist.

Additionally, human variations within a trainee may color a training event, so that outcomes on identical tasks are not consistent for a given trainee over time. These dynamic intra-personal factors may change from moment-to-moment and are often classified as state-based variables. For example, a person's ability to focus on a task rarely remains stable over extended periods of time, instead fluctuations from highly engaged to extreme bored are normal. The mix of inter-personal and intra-personal differences prohibits a 'one size fits all' approach to training system design. Therefore, instructional designs must include adaptable features that afford individualized tailoring to maximize each training events for each trainee.

The second component, the instructor, is the primary resource currently available for providing personalized tailored training. In a one-to-one context an instructor can reasonably estimate a trainee's state and make real-time adjustments to better match instruction to a trainee's current needs in real-time within a training context. Unfortunately, the one-to-one approach is not cost a cost effective method, nor is it reasonable from a manpower perspective; skilled personnel/instructors are a valuable resource. A more common approach is to have a single instructor provide CBT to multiple trainees at a time, thus, diminishing the value of the highly skilled instructor. Monitoring and managing multiple trainees overwhelms an instructor's mental bandwidth making it nearly impossible to estimate a trainee's current needs and make instructionally significant adjustments. Moreover, the high workload requirements placed on instructors becomes more troubling for novice instructors that lack the abilities gained through experience.

The final component, the instructional technology, is the mediating component that has the greatest potential to enhance, or disrupt, the training environment. In a CBT context, the technology (hardware and software) connects both the trainee and instructor to the instructional materials and it connects them to each other. Thus, favorable outcomes are directly related to the quality of the human system integration. Unfortunately, this important point is often overlooked because the objectives of a technologist are derived from goals that are indirectly linked to the basic premise of instruction, impart knowledge and skills. The disconnect results in an increased risk for the development of ineffective instructional tools. For example, improvements in 3D graphics may improve collaboration work, but this is of little value to an instructor that is training KSAs for a one-person task. Moreover, instructional technologies with form factors that don't match the training environment are not good candidates for transition to field settings. Therefore, a human-systems integration philosophy is particularly useful for the design, development, and implementation of advanced instructional concepts and technology.

2.1 Human Performance

Over the past one hundred years, models of human performance and behavior have indicated that KSAs tend to fit within scoped boundaries, such as, the inverted U-hypothesis [1], zone of proximal development [2], comfort zone [3], and flow state [4]. The overarching take-away from the extant literature is that there appears to be a "sweet spot" for getting people to perform at their best! While it appears that we all

have a sweet spot it is also clear that finding the sweet spot is highly individual. What stimulates the flow state in one person may frustrate, or bore, someone else. Moreover, it is not easy to get into or maintain this highly desirable state of optimal performance. There are techniques that can help initiate flow; however, they require meta-cognitive skills, such as, self monitoring and regulation of cognitive states. Unfortunately, timely attainment of these higher order skills is likely beyond the reach of most people. These are implicit skills that are difficult to quantify and are not readily observable. However, the field of augmented cognition is making strides toward the design of technology/tools that may make these implicit processes explicit. In the near-term, these technologies may provide insight to instructors about a trainee's state that may enable for the design of systematic methods to stimulate a flow state. An instructional design that integrates the human-technology components so tightly has great potential to deliver personalized instruction in a number of new and interesting ways.

❖ *But lo! Men have become the tools of their tools*, Thoreau, Henry David

In 1854, Henry David Thoreau raised an idea (maybe concern) to emphasize the inter-relation between people and technology. Leastwise, his statement illustrates how technology "tools" are more than mere objects; they are integral implements that interact with people to facilitate achievement of goals. Often, advanced technologies are not accepted by military leaders and/or instructors because they are too disruptive, either in concept of operations or technical execution. But, also, there is this idea of the status quo in which people are comfortable using current tools in the manner to which they have grown accustom. Whether, or not, the new tool is better suited for the task at hand is moot if it is judged to be too farfetched. To balance the acceptance-advanced equation, the initial Trainee Engagement Management System (TEMS) solution is being designed as an enhancement to a government owned Instructor Operator Station (IOS) that aligns with practices currently employed by military instructors. In short, introducing a simple design modification to an existing CBT context might be an acceptable application of advanced physiological technologies.

2.2 Instructional Quality

In broad terms, instructional quality can be tailored through monitoring and management of two types of variables - situation and person. Situation variables are the external factors and conditions of an instructional context that frames learning events. Because CBT platforms afford relatively easy monitoring and management of situation variables, it continues to be a very popular and useful option for the delivery of many instructional methods, such as, demonstration based training and simulation based training. Conversely, there remains a weakness in the instructional administration of CBT for the direct observation and management of person variables, such as cognition, affect, and attention that influence instructional quality. Maintaining engagement, or focused attention, is a very important person variable for knowledge and skill acquisition. Finding ways to stay focused is challenging for everyone. Our lives are filled with many things that compete for our attention, creating distraction and

depleting our mental resources. In the case of military personnel that are immersed in a culture of always on technology it is common for them to lack the mental energy required to maintain focused attention during CBT. If a trainee is not engaged in the CBT event the path to learning is blocked impairing both training effectiveness and efficiency. The approach described in this paper acknowledges that trainee engagement is a cornerstone to the development of high performers. Specifically, comprehension and retention are affected by the trainee's degree of active participation (i.e., effort) in the training event. Paas [5] describes how the path to expertise depends on an individual's willingness to work at becoming expert:

> ❖ *In research on deliberate practice, it has been noted that because this type of practice requires trainees to stretch themselves to a higher level of performance, it requires full concentration and is effortful to maintain. This does not make it a very enjoyable experience, so without the motivation to improve, trainees will soon give up (Ericsson et al., 1993). Feedback can play a crucial role in their willingness to continue to invest effort (indeed, feedback is also considered to play a crucial role in deliberate practice; see, e.g., Ericsson & Lehmann, 1996). [6]*

Ideally, each trainee would be drive by an intrinsic motivation to achieve expert status and, to that end, would supply the effort required to maintain engagement during a given training event. However, that ideal circumstance is not typical of trainee populations in real world contexts, nor is it reasonable to expect that any individual could consistently sustain that level of motivation. Thus, instructional designs that systematically foster active learning in typical training environments are desirable.

3 Instructional Design Concept

The TEMS design concept offers an augmented cognition solution to close the person variable loop. Augmented cognition is a field of research that continues to re-design Human Computer Interaction (HCI), as it makes technological systems responsive to state-based person variables. An opportunity exists to exploit Commercial Off-The-Shelf (COTS) technologies that assess physiological arousal via measures of Electrodermal Activity (EDA) to improve instructional effectiveness and efficiency of CBT approaches. A new class of EDA technologies has successfully transitioned from controlled laboratories to real world settings. With improved form factors and low cost, these technologies are good candidates for near-term advancements in training contexts. Reliably, EDA measures arousal and can provide indications of high arousal (e.g., excitement, engagement, anger) and low arousal (e.g., boredom, disengagement, calm) that affect cognition and emotion. We have conceptualized an instructional design concept that utilizes EDA to mitigate information loss in a CBT context

Trainee engagement is a critical component for optimizing instructional effectiveness and efficiency of CBT. Often, a single instructor simultaneously administers CBT exercises to multiple trainees making it difficult for them to detect every

occurrence of poor engagement and/or motivation. Thus, they are unable to provide crucial feedback or prompts that would re-engage the trainee. A key capability of TEMS is reporting to instructors those trainees lacking proper engagement during training exercises [6]. The TEMS concept focuses on the management of trainee engagement to maximize learning opportunities. To accomplish that we conceptualized a design that enhances an instructor's awareness of trainees' state to augment their judgments and improve decision making.

At one location, an instructor is able to monitor both behavioral performance and physiological indicators of a trainee's arousal. These two pieces of information combine to convey a more complete picture of a trainee's performance, or progress. In many instructor operator stations (IOSs) the instructor has the option to peer into a trainee's lesson so they can observer trainee performance. However, the ability to peer into a trainee's lesson only provides observations of situational variables – a very limited assessment of performance as discussed earlier in this paper. Paring a physiological assessment of arousal with the standard capability to observe trainee behaviors will augment instructor's ability to provide instructional support.

From a human-system perspective, the core capabilities of the TEMS instructional design concept are embodied in the IOS. The IOS contains the key systems features that enable the advanced training capabilities of the TEMS design concept. Moreover, the IOS is the system interface to the intelligence that drives the technology, the instructor. At the IOS, an instructor is able to monitor trainees' behaviors (performance) and physiological states (processes). The enhanced monitoring feature within the IOS enables the deployment of re-engagement strategies promoting a transactional model for instructional design, see Figure 2 below.

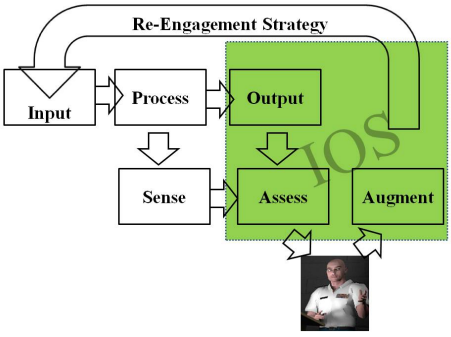


Fig. 2. Transactional Model for Instructional Technology

The transactional model for instructional design incorporates two process models: 1) Information-Process Output (IPO) for human performance, and 2) sense-assess-augment for human performance augmentation. The top three IPO boxes, in Figure 2, are conceptualized as follows; all stimuli in the world are symbolized as 'input', cognitive and biological processes are 'process', and behavioral actions taken are represented as 'outcome'. The United States Air Force's Human Performance Augmentation model is represented in the second row of 3 boxes and is conceptualized as;

sensor monitoring provides 'sense', bio-data classification acts as 'assess', and intervention strategies assume the role of 'augment'[7]. As a foundation for TEMS, these models are used to characterize trainee performance so that the interplay between trainee behavioral performance and physiological arousal are exposed to instructors to enhance their awareness via an IOS.

Note, the instructor is in the loop and is responsible for deploying re-engagement content to the trainees. The prototype system will include an empirically supported metacognitive prompting strategy as its re-engagement strategy. Many opportunities exist to be creative with respect to designing re-engagement strategies and much research is needed to clearly link strategies to specific types of trainee needs. However, recent work suggests that the State-based Information-loss Process (SIP) Model could be used to appropriately administer a metacognitive prompt strategy.

3.1 Advanced Learning Concepts

The SIP Model was used to inform the initial prototype of the TEMS design concept. The SIP Model [8] identifies possible points for cognitive breakdown that may contribute to information loss during instruction. It is an evidence-supported model that focuses on higher-order learning (i.e., knowledge/skill integration and application) that is important for imparting complex KSAs that are required in military operations. Based on the SIP Model, we incorporated a metacognitive prompting strategy that could be used to mitigate information loss in a CBT context.

❖ *Learning without thought is labor lost, Confucius*

Metacognition is widely accepted as an implicit "thinking" skill that enhances one's ability to learn and solve problems [9]. Recent findings from the Next-generation Expeditionary Warfare-Intelligent Training (NEW-IT) program have demonstrated increases in learning effectiveness when metacognitive prompts were employed in the service of learning [10]. Those successes demonstrate that these prompts can be effective in a CBT context. We will build on those findings to develop a system that will help instructors more precisely target trainees for instructional intervention, metacognitive prompts. Leveraging a COTS state-based assessment of trainee arousal (i.e., EDA) the TEMS concept re-designs how this validated instructional strategy could be implemented. The physiological-based aspect of TEMS provides an objective assessment a person variable that allows instructors to more effectively, efficiently, and confidently employ an advanced strategy with a familiar CBT context. In addition to a meta-cognitive prompting strategy that we plan to test in the initial instantiation of the TEMS design concept, other prompting strategies could easily be adapted. Paas [5] outlines a few candidate prompting strategies that may be employed to promote deep comprehension and self regulation:

- Reflection prompts: to promote self-regulated learning competency and sustainability through reflection on one's own learning
- Self-explanation prompts: to promote understanding of the underlying principles of a problem, often provided with worked examples

- Critical thinking prompts: to promote the abilities to evaluate one's own thinking and fill in gaps in knowledge

Incorporating and combining empirically validated learning techniques that match the learning needs of each trainee is fundamental to the overarching approach for the TEMS design concept. Our objective is to bring the best practices of instruction forward by design and to do so by integrating precision assessments of each trainee's engagement via behavioral and physiological assessments.

The strength of the approach comes from a convergence of advanced technologies and best practices from a variety of disciplines. However, that means we will have to tackle many technical, theoretical, and practical challenges to realize a fully implemented system. Among the priorities, is to validate best practices/guidelines that identify which instructional prompts are best suited for a topic, a trainee's scaled competency, and/or learner's style and state. While some guidance exists in for the use of instructional strategies there remains a gap for implementation in a state-based system.

4 Envisioned System

The TEMS design concept is an innovative instructional design that supports instructors in the management of trainee engagement. Among other objectives, the work described in this paper is attempting to create a field ready system that simultaneously introduces both advanced technology and learning strategies. The TEMS design concept is envisioned as a domain agnostic instructional technology that could be used to support any instructor in achieving their goals, deliver high quality instruction to as many trainees as possible. To support those objectives, we have conceptualized our design concept within a common CBT context that shares many features with a traditional classroom setting, see Figure 3.

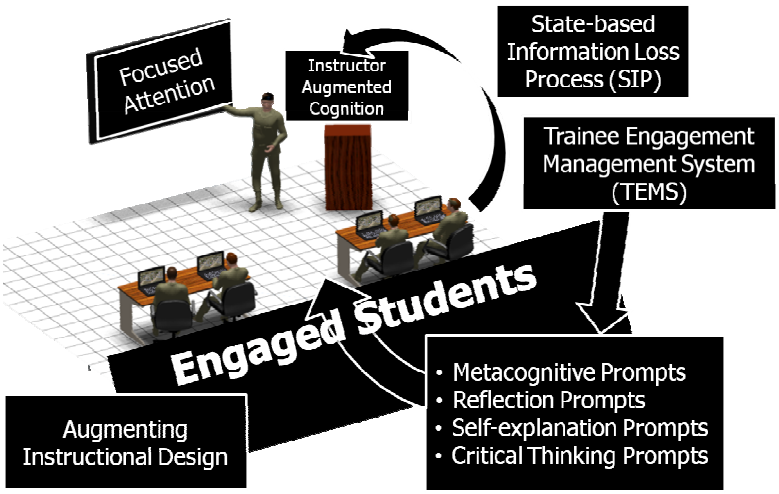


Fig. 3. Traditional Classroom Features

The basic idea is to provide instructors real-time assessments of engagement for each trainee in the instructional setting. Visualizations of engagement will be displayed on an Instructor Operator Station (IOS) allowing instructors to get a top-level view of trainee engagement at a single location. When a trainee's engagement drops below an acceptable level, the instructor can easily identify which trainee is not properly engaged. When a state-based change is observed the instructor may choose to take steps to re-engage that trainee or they may decide see if the trainee self-corrects to regain engagement. The TEMS design concept embraces a human systems integration philosophy that supports, not replaces, instructors to promote tailored instruction in CBT environments.

5 Discussion

A mix of technical and technological change throughout the military contexts imposes a great deal of responsibility on military personnel to reach and maintain mission readiness. Effectively and efficiently imparting mission-critical KSAs is a first order goal of instructors across the military services. Under ideal one-on-one conditions it can be difficult to optimize that match instructional content and a trainee's unique needs. Individual differences and state-based variations converge at each training session to create a signature experience. Thus, delivery of high quality instruction at each training session requires an adaptive capability that is responsive to the dynamic person variables. To date one-on-one paradigms with highly skilled instructors are the best way to consistently obtain that quality of instruction. However, that paradigm is extremely costly and is not sustainable, or reasonably feasible, in military training settings. Typically, trainee settings include one instructor that is responsible for a large number of trainees with a ratio that is closer to 1:10, as compared to 1:1; thus, a need for innovative instructional designs persists.

A significant amount of work continues to be devoted to the development of Intelligent Tutoring (IT) systems that can be used to support training in unsupervised contexts and replace live instructors. The explicit goal of IT is to achieve the same level of quality as observed in one-on-one instruction. However, this is not likely to be realized in the near-term as a field-ready solution. In large part, much work remains in the fields of Artificial Intelligence (AI) and Natural Language Processing (NLP) before this worthy goal can be validated and transitioned. For that reason, the AI capabilities required to support the TEMS design concept are comparatively crude. The TEMS design concept only provides indications of engagement that can be derived from commercially available data classifiers. The actual intelligence within the TEMS concept resides within the instructor's expert judgment and highly skilled decision making. As the research community demonstrates advances in AI capabilities future implementations will incorporate intelligent tutoring capabilities; however, the envisioned system is intended to always include roles and responsibilities for a live human instructor.

References

1. Yerkes, R.M., Dodson, J.D.: The Relation of Strength of Stimulus to Rapidity of Habit-formation. Journal of Comparative Neurology and Psychology 18, 459–482 (1908), http://psychclassics.yorku.ca/Yerkes/Law/
2. Vygotsky, L.S., Whorf, B.L., Wittgenstein, L., Fromm, E.: Language and Consciousness. In: Pickering, John, Skinner, Martin (eds.) From Sentience to Symbols: Readings on Consciousness, Harvester Wheatsheaf (1990)
3. Hancock, P.A., Warm, J.S.: A Dynamic Model of Stress and Sustained Attention. Human Factors 31(5), 519–537 (1989)
4. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper Perennial, New York (1991)
5. Paas, F., van Gog, T.: Principles for Designing Effective and Efficient Training of Complex Cognitive Skills (2009), doi:10.1518/155723409X448053
6. Craven, P.L., Tremoulet, P.D., Barton, J.H., Tourville, S.J., Dahan-Marks, Y.: Evaluating training with cognitive state sensing technology. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) Augmented Cognition, HCII 2009. LNCS, vol. 5638, pp. 585–594. Springer, Heidelberg (2009)
7. Blackhurst, J.L., Gresham, J.S., Stone, M.O.: The Quantified Warrior. Armed Forces Journal (2012), http://www.armedforcesjournal.com/2012/12/12187387 (retrieved)
8. Vogel-Walcutt, J.J., Bowers, C.A., Marino-Carper, T., Nicholson, D.: Increasing Learning Efficiency in Military Learning: Combining Efficiency and Deep Learning Theories. Military Psychology (2010)
9. Brown, A.: Metacognition, Executive Control, Self-Regulation, and Other more Mysterious Mechanisms. In: Reiner, F., Kluwe, R. (eds.) Metacognition, Motivation, and Understanding, pp. 65–116. Erlbaum, Hillsdale (1987)
10. Lackey, S.: Next-Generation Expeditionary Warfare Intelligent Training (NEW-IT) Program Summary Booklet (2011), http://active.ist.ucf.edu/LinkClick.aspx?fileticket=K0%2bcwKl0brQ%3d&tabid=430

Instrumenting Competition-Based Exercises to Evaluate Cyber Defender Situation Awareness

Theodore Reed, Kevin Nauer, and Austin Silva

Sandia National Laboratories, Albuquerque, NM, USA
{tmreed,ksnauer,aussilv}@sandia.gov

Abstract. Cyber defense exercises create simulated attack and defense scenarios used to train and evaluate incident responders. The most pervasive form of competition-based exercise is comprised of jeopardy-style challenges, which compliment a fictional cyber-security event. Multiple competitions were instrumented to collect usage statistics on a per-challenge basis. The competitions use researcher-developed challenges containing over twenty attack techniques, which generate forensic evidence and observable second-order effects. The following observations were made: (1) a group of defenders performs better than an individual; (2) situation awareness of the fictional event may be measured; (3) challenge complexity does not imply difficulty. This research introduces a novel application of system instrumentation on competition-based exercises and describes an exercise development methodology for effective challenge and competition creation. Effective challenges correctly represent difficulty and reward competitors with objective points and optional forensic clues. Effective competitions compliment training goals and appropriately improve the knowledge and skill of a competitor.

1 Introduction

Information (cyber) security exercises have become powerful tools for simulating and planning for emergency scenarios, training, and competition. This paper focuses on the latter examples of training and competition. These exercises create simulated attack and defense scenarios where participants organize into groups and interact hands-on with operating systems, hardware, and software.

The exercise format varies, including modes with a sizable red (or attack) team versus many blue (or defending) teams, all red versus red teams, or all blue versus blue [CA1]. The red versus red is considered an attack and defense exercise where each team functions as both blue and red; they must maintain their security posture while decreasing their opponent's. The red versus blue is an interactive defense where each blue team is evaluated by their security posture after a complex and distributed set of red team attacks. A blue versus blue exercise uses point-valued challenges; the team that correctly solves the most challenges is the exercise victor [DE1].

The blue versus blue, or challenge-based exercises, are well-suited for training. The instructor develops interactive-challenges (i.e., a capture of forensic

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 80–89, 2013.

data containing a reportable sliver of evidence) which requires comprehension of course material to solve. Students may be motived to learn the material such that they can demonstrate competitive mastery (we do not make this assertion).

Challenge-based exercises are also the most flexible. Participants typically use their own hardware and tools, and may compete remotely and asynchronously (e.g., an exercise may not be bounded by time). Unfortunately this flexibility creates a difficulty for instrumentation; it is difficult to observe behavior and interaction. In this paper we describe a methodology for competition-based exercise development that yields measurable usage data and allows competition-designers introspection into player-challenge interaction.

1.1 Purpose of Study

Competition-based, continuous [GG1], exercises have proven successful for multiple applications and have become a pervasive [CB1] method of comprehension verification and community entertainment. Similar formatted exercises have been commonplace in high consequence domains (e.g., military) [MT1]. However, there have been few studies on the development and operation of these exercises and the human interaction in the cyber-security domain.

This research introduces an exercise platform and challenge development methodology that allows study of player-exercise, and player-player interaction. Example studies include: (1) a comparison of training modes; (2) player and tool adaptability; (3) situation awareness comprehension variability [T1]; (4) defensive solution-path discovery [SH1]; and (5) challenge playability tolerance. The last example uses the exercise to collect interaction statistics and create an arbitrary game mechanic called tolerance [GD1]. This demonstrates the exercise platforms ability to verify the challenge development, and is part of the development methodology. The methodology defines four categories of tolerance: simple, difficult, confusing, and unsolvable. A well-defined challenge should both be simple or difficult, and generate measurable feedback effects.

2 Approach

2.1 Exercise Platform

This research used a jeopardy–style interface containing categories of increasing-value challenges to represent the exercise. This game–board uses username and password account (or user) authentication and associates each user to a team. If any user correctly solves a challenge the team will receive the point-value; a team score is the aggregate of its users. The interface presents a robust configuration to the competition-designer.

The designer chooses from an XML-defined repository of challenges, with the ability to set time-thresholds and custom point values for each. A challenge is

defined as a block of instruction, suggested time to complete, suggested point value, and solution. A solution may be an input string, a review process, or a trigger event. These challenges are organized into categories and categories are organized into boards. The designer configures the board availability (i.e., start and stop time) as well as trigger events (i.e., stop conditions) and submission rules. Fig. 1 shows an example participant view of the game–board. Note that one 100-point challenge has been solved by the user.

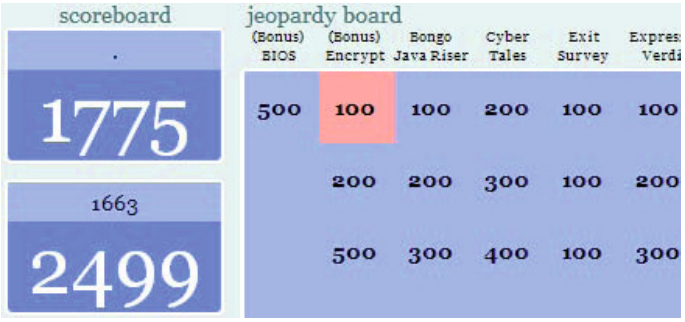


Fig. 1. Game–board from a participant's view

2.2 Methodology

Developing exercise challenges is non-trivial. Challenges should test a participant's critical thinking and knowledge application abilities. Challenges should implement a 1:1:2 ratio of effort required for a solution. This ratio represents 1-part discovery, 1-part understanding, and 2-parts solution development. The participant should spend the discovery phase analyzing the challenge to find a starting point. The understanding phase should be spent researching what skills, tools, and techniques are required for a solution. The solution development should stress the participants technical and critical-thinking prowess.

The challenge developer must maintain the highest level of fidelity for their challenge. Environment and data anomalies jeopardize the tolerability of a challenge and degrade any potential experiment or assessment. Example anomalies may include: (1) improper use of IP-space when creating a synthetic environment for forensic data generation, (2) unmatched operating system version artifacts left in physical memory, (3) poorly synchronized timing seen in network data, file systems, and descriptions, and (4) typographic fixes or incorrect checksums.

Dependent Challenge. A challenge (*c1*) may include artifact data needed to solve a separate challenge (*c2*). Challenge *c2* is called a dependent challenge. Dependent challenge development is particularly difficult; the development must be conscience of the playability implied by lack of depended knowledge.

2.3 Participants and Data

This research used five exercises. Each spanned at least two working-hour days, comprised of the same challenge set and over 220 combined participants. The participants represent a combination of high school students, undergraduate and graduate college students, and industry professionals. There were a total of 95 teams with a majority of 1-player teams with an assumed[1] maximum of 7-player teams. For this research no identifiable information was collected. When each exercise is completed usage data is exported with teams and users represented as arbitrary integer placeholders.

The exercises used 97 challenges per-event. Challenges were worth 100-500 points each, and most were solvable independent of the others. In all of the exercises recorded, wrong answers had no penalty and awarded 0 points; challenges were attempted until a successful submission (if any). The exercises attempted to measure participant situation awareness about a fictional cyber-security event. The challenges contained forensics data which required little interaction with the exercise platform. Thus it was very important that the challenges generate second-order effects such as (1) red-herring[2] submissions, (2) fictional names, services, or IP-addresses, or (3) additional forensics data.

The analysis uses an example assessment of challenge playability tolerance. Submissions and incorrect actions are compared to create a tolerance. Over six thousand submissions were recorded with just fewer than one thousand correct submissions. Over one million actions were recorded with a ratio of 4:1 incorrect to correct actions per challenge.

3 Results

3.1 Data Sanity

The data from all five exercises is combined and visualized in the following sections. In Fig. 2 the total score for each team is plotted in ascending order. The score distribution follows the exponential trend-line very closely. This is expected as better-performing teams solve higher-valued challenges across all categories. Problems with challenge confusion, which require participants to guess, may create a deviation. An imbalance in scores is highlighted indicating a potential guessing situation.

In Fig. 3 the number of correct and incorrect submissions per-challenge are plotted with a logarithmic trend-line. This describes a global interaction for every challenge. Challenge developers should expect a global logarithmic distribution, indicating a well-formed exercise with increasingly difficult challenges.

[1] One of the exercises was played virtually, thus any team may contain an unknown number of human players whom share user accounts. However, it is unlikely that accounts are shared as the game platform does not allow simultaneous challenge solving (i.e., only one challenge can be viewed at a time).

[2] A known-wrong submission that is easy or obvious but indicates progress.

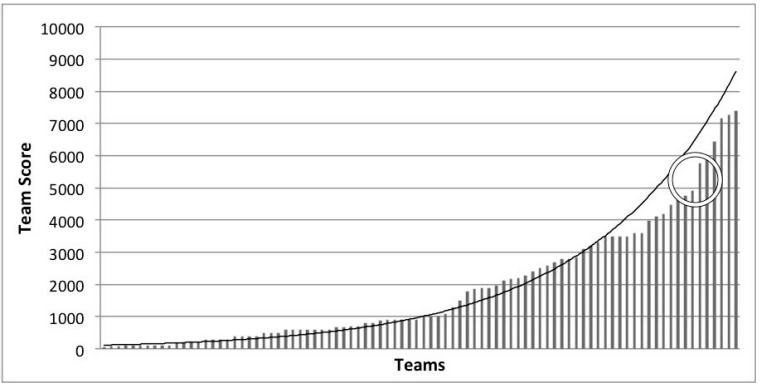


Fig. 2. Total score distribution with exponential trend-line

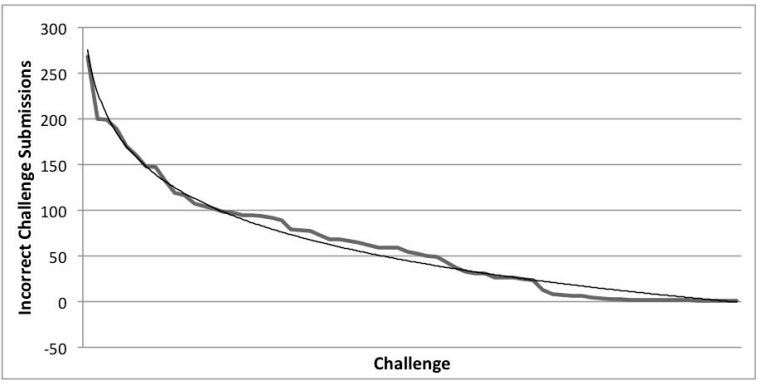


Fig. 3. Submission count per-challenge with logarithmic trend-line

Any abnormalities or deviations in submission counts may indicate confusion. The ratio of incorrect to correct submissions in Fig. 9 is used to enhance this visualization and help identify poorly-defined challenges. The two highlighted challenges have enormous incorrect to correct submission ratios.

Using a linear-trend with a bisection creates four quadrants of challenge ratios. Challenges with high submissions and low correct submissions (Q1) are candidates for review. Balanced ratios with high submissions are also candidates if not defined as difficult challenges. The same applies to the inverse if not defined as simple challenges. Finally, challenges without correct submissions are flagged as potentially unsolvable[3].

[3] Occasionally a 'solvable-but-near-impossible' challenge is useful for attracting curiosity.

3.2 Activity

In Fig. 4 the average momentum for the top three teams overall is shown as the dark line. The average momentum for the top three teams from one standard deviation (sd) away is shown as the light line. The momentum is seemingly linear for both groups. In this representation where momentum is a function of score versus time the reason for a dramatic (30%) point spread is unknown.

The point spread is more obvious when comparing Fig. 5 and 6. These figures show a normalized delay between submissions for each set of three teams. The longer each team plays, the more frequently they experience delayed submissions. Note, this does not represent periods of non-play. Delay normalization is a function of incorrect submissions. These plots may suggest teams are encountering more difficult challenges. The plots corroborate a similar momentum in Fig. 4 with a similar delay from point 15.

3.3 Tolerance

In Fig. 7 and 8 participant tolerance is show as the average for the top three teams and the average for the top three teams from one *sd*. To assess tolerance the exercise platform measures a combination of player frustration (f) and promotion (p). A promotion p, is defined as any positive feedback provided by the exercise platform to the player. A frustration f is a continually increasing value assigned to each player; f is reset to an initial state upon p. The exercise platform measures team frustration using a gain calculation based on incorrect actions and time.

$$f_t = \sum_{i=p}^{n} n\,(t_i - t_{i-1}) \qquad \text{where } p \text{ is the last promotion event } . \tag{1}$$

Fig. 8 shows a significant amount of frustration toward the end of the measurement which most likely leads the disparity in points. Within the five exercises a p is a correct submission or a positive action taken by a participant (i.e., acquiring an additional piece of forensics data, gaining access to a services, or disabling an attacker).

3.4 Situation Awareness

Situation awareness is assessed by comparing the average performance of event related challenges to non-event related challenges. The event related challenges implicitly include artifacts and relations to other event challenges. These relations are not dependent challenges; the related challenges are solvable independently. However, knowledge of additional event related challenges builds context around possible attack vectors, techniques, and tools. If the participant has situation awareness and can build this context, the assertion is they will solve event related challenges more efficiently.

Out of the 97 challenges, 13 tightly related challenges were compared against an unrelated 13. These pairs were assessed by the challenge developers as having

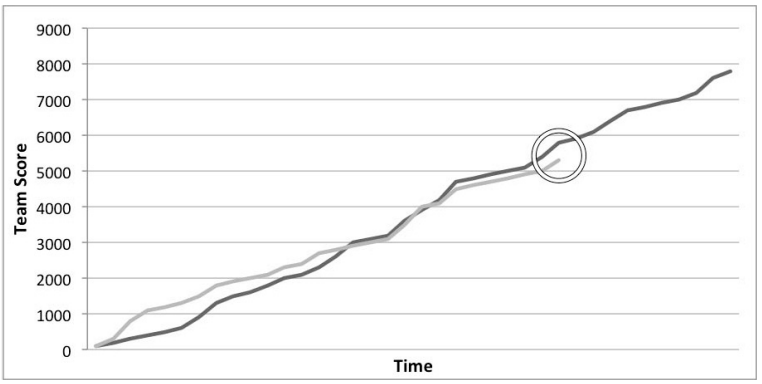


Fig. 4. Average momentum of top three teams (dark) and top three teams from one standard deviation (*sd*)

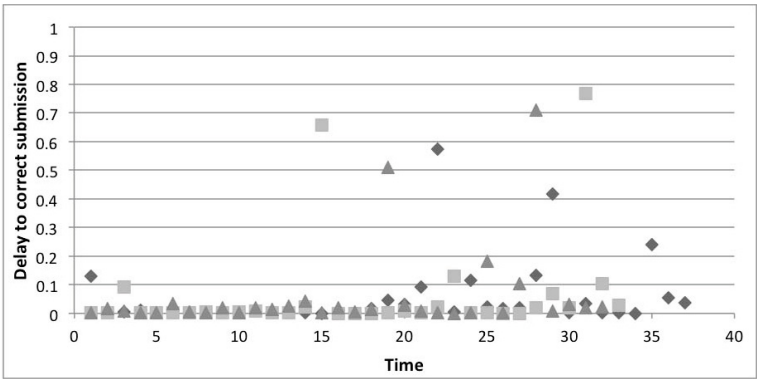


Fig. 5. Normalized time delay between submissions for top three teams

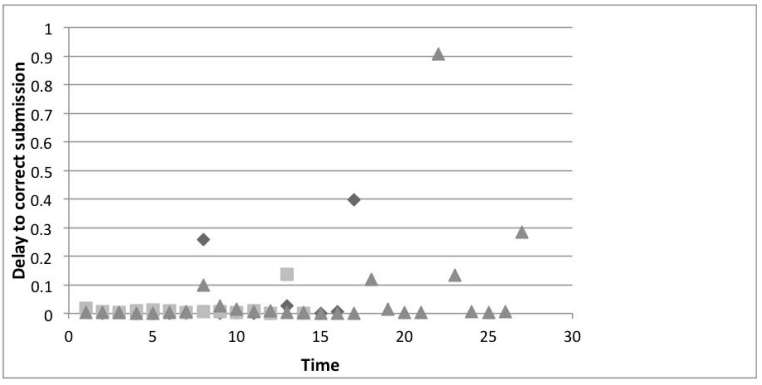


Fig. 6. Normalized time delay between submission for top three teams from one *sd*

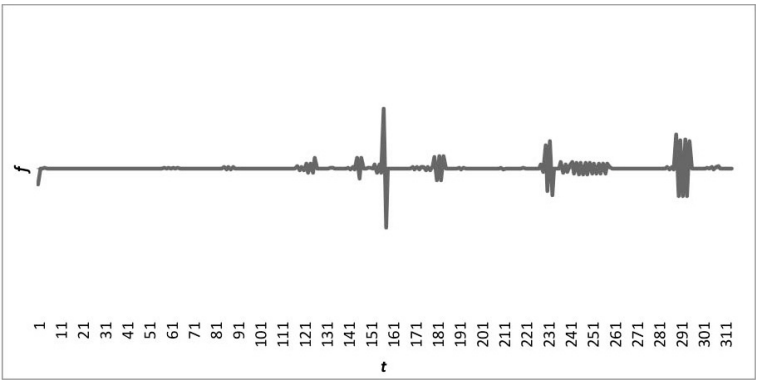


Fig. 7. Average frustration for top three teams

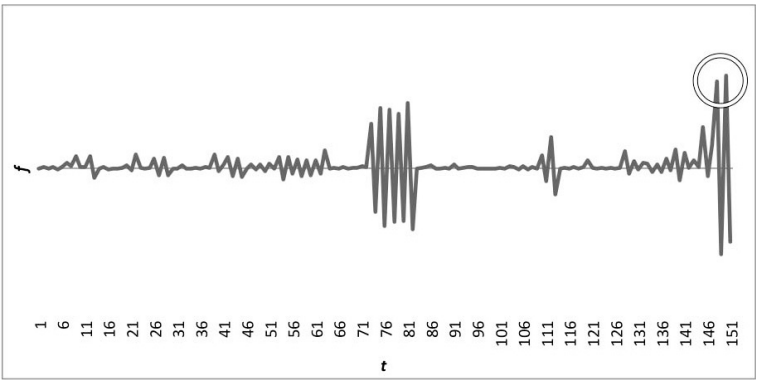


Fig. 8. Average frustration for top three teams from one *sd*

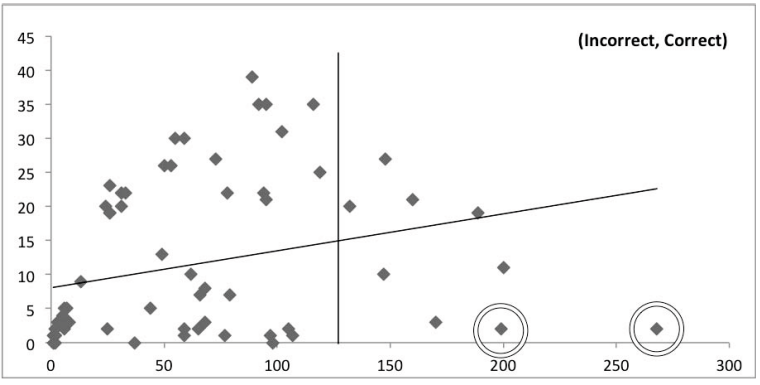


Fig. 9. Challenge submission ratios plotted (incorrect versus correct)

similar difficulty with little knowledge overlap. It is possible to solve the pairs in either order (non-event then event, or event then non-event) without confounding performance. Over 96% of participants demonstrate a better average performance on the event related challenges. A degree of situation comprehension may be measured within the exercise using a tool called *Plotweaver*[O1].

4 Conclusion

This exercise platform successfully demonstrates an example evaluation of cyber defender situation awareness. Participants are evaluated by their comprehension of a fictional cyber event through narration and plot description. Participants are also evaluated based on event related challenge performance versus non-event related challenges. If both a related and non-related challenge exists with similar difficulty and no overlap in knowledge requirement or other confounds: then the solution path can evaluated based on insight. The platform generates these statistics by comparing measurements generated through instrumented challenges.

The platform successfully validates challenge tolerance through usage statistics. This feedback is given to challenge developers and functions to remove unwanted difficulty confounds. Challenges that move from unsolvable or confusing to difficult make the exercise more enjoyable, reduce potentially harmful frustration, and generate more statistically-relevant usage data.

Instrumentation of challenges to provide measurable second-order effects created observations on player activity fallout based on frustration thresholds. This activity was not apparent in objective interaction data such as game-board activity and score momentum.

5 Future Work

Additional objective and subjective usage measures will continue to enhance the community's ability to use cyber defense exercises to improve domain knowledge and event response. The existing measures can be engineered into the exercise platform to provide real-time feedback to the designer. If player frustration and interaction threshold classes can be defined, a designer can provide in-line challenge and exercise augmentations. These augmentations can reduce frustration and experiment confounds to generate better data and a more enjoyable exercise experience.

Finally, challenge solution paths should be more closely monitored. Additional rewards can be granted to players demonstrating unique solutions. This encouragement may potentially enhance situation awareness, generate richer usage data, and reduce future frustration thresholds.

References

[T1] Tadda, G.P.: Measuring performance of Cyber situation awareness systems. In: Proceedings of the 11th International Conference on Information Fusion. Rome Res. Site, Air Force Res. Lab., Rome, NY, pp. 1–8 (2008)

[GG1] Glicksberg, I., Gross, O.: Notes on Games over the Square. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games. Annals of Mathematics Studies 28, vol. II, pp. 173–183. Princeton University Press (1950)

[GD1] Gilleade, K., Dix, A.: Using frustration in the design of adaptive videogames. In: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE 2004), pp. 228–232. ACM, New York (2004)

[O1] Ogievetsky, V.: PlotWeaver (2013), `https://graphics.stanford.edu/wikis/cs448b-09-fall/FP-OgievetskyVadim`

[MT1] Mullins, B., Lacey, T., Mills, R., Trechter, J., Bass, S.: How the Cyber Defense Exercise Shaped an Information-Assurance Curriculum. In: IEEE Symposium on Security and Privacy, pp. 40–49 (2007)

[CB1] Childers, N., Boe, B., Cavallaro, L., Cavedon, L., Cova, M., Egele, M., Vigna, G.: Organizing large scale hacking competitions. In: Kreibich, C., Jahnke, M. (eds.) DIMVA 2010. LNCS, vol. 6201, pp. 132–152. Springer, Heidelberg (2010)

[DE1] Doup, A., Egele, M., Caillat, B., Stringhini, G., Yakin, G., Zand, A., Cavedon, L., Vigna, G.: Hit 'em where it hurts: a live security exercise on cyber situational awareness. In: Proceedings of the 27th Annual Computer Security Applications Conference (ACSAC 2011), pp. 51–61. ACM, New York (2011)

[CA1] Cowan, C., Arnold, S., Beattie, S., Wright, C., Viega, J.: Defcon Capture the Flag: defending vulnerable code from intense attack. In: Proceedings of the DARPA Information Survivability Conference and Exposition (2003)

[SH1] Sommestad, T., Hallberg, J.: Cyber Security Exercises and Competitions as a Platform for Cyber Security Experiments. In: Jøsang, A., Carlsson, B. (eds.) NordSec 2012. LNCS, vol. 7617, pp. 47–60. Springer, Heidelberg (2012)

Enhanced Training for Cyber Situational Awareness

Susan Stevens-Adams, Armida Carbajal, Austin Silva, Kevin Nauer, Benjamin Anderson, Theodore Reed, and Chris Forsythe

Sandia National Laboratories, Albuquerque, NM, USA
{smsteve,ajcarba,aussilv,ksnauer,brander, tmreed,jcforsy}@sandia.gov

Abstract. A study was conducted in which participants received either tool-based or narrative-based training and then completed challenges associated with network security threats. Three teams were formed: (1) Tool-Based, for which each participant received tool-based training; (2) Narrative-Based, for which each participant received narrative-based training and (3) Combined, for which three participants received tool-based training and two received narrative-based training. Results showed that the Narrative-Based team recognized the spatial-temporal relationship between events and constructed a timeline that was a reasonable approximation of ground truth. In contrast, the Combined team produced a linear sequence of events that did not encompass the relationships between different adversaries. Finally, the Tool-Based team demonstrated little appreciation of either the spatial or temporal relationships between events. These findings suggest that participants receiving Narrative-Based training were able to use the software tools in a way that allowed them to gain a greater level of situation awareness.

Keywords: cyber security, training, situational awareness.

1 Introduction

Situation awareness is essential to effective cyber security analysis and incident response team performance. However, cyber situation awareness has not been well studied (Tadda, 2008). This research sought to help clarify the cyber situation awareness problem, while providing insights that will improve training effectiveness for cyber defenders.

An explosion of new vendor and open source tools has occurred in the past few years to address the growing cyber problem, with U.S. Government enterprise networks and their incident response teams being a primary market. However, these new tools have not always improved the situation awareness of cyber security analysts. Consequently the return on investment has been questionable given the costs of purchase, development and integration of the new technologies.

Nonetheless, cyber security analysts need tools to assist them in fathoming the vast quantities of data and deciphering ever-more sophisticated network attacks. There is

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 90–99, 2013.

need for research to understand why tools that ought to increase the productivity of cyber security analysts often fail to realize this objective. We believe that this failure may be partially attributable to insufficient training and, particularly, the fact that intended users often lack fundamental knowledge essential to effectively use the tools being provided to them. Today, there is no scientific basis for asserting that one mode of training cyber defenders to use software tools is superior to any other mode of training. Likewise, there has been no openly published empirical assessment of students receiving alternative modes of training. The objective of this project was not to compare alternative software tools and no data was collected that reflected on the relative performance or utility of alternative software tools. Instead, through laboratory research employing human performance measurement, the current project scientifically addressed the question of what type of training is needed to maximize the effectiveness of new tools being introduced to improve the situation awareness of cyber security analysts.

1.1 Purpose of Study

The current project employed a suite of network analysis tools comparable to those commonly used in operational cyber settings. Two modes of training were considered. The baseline training condition (Tool-Based training) was based on current practices where classroom instruction focused on reviewing the software functionality with various exercises in which students apply those functions. In the second training condition (Narrative-Based training), classroom instruction addressed software functions, but in the context of adversary tactics and techniques. Upon completion of training, participants were evaluated during a Tracer FIRE (Forensic and Incident Response Exercise) simulated blue team exercise. It was hypothesized that students receiving Narrative-Based training would gain a deeper conceptual understanding of the software tools and that this would be reflected in better performance during the Tracer FIRE exercise.

Three hypotheses were tested. Hypothesis 1: The narrative-based training is different from the tool-based training and will result in better performance in an assessment of students' abilities to use software tools to interpret events associated with a cyber-attack. Hypothesis 2: Personality has an effect on team success and dynamics. Certain personality attributes will result in lower team scores. Hypothesis 3: Cognitive aptitude has an effect on team success. Certain cognitive aptitudes will result in superior team scores.

While research of this nature is commonplace in other high consequence domains (e.g., military operations), there exists little precedent within the cyber security domain. Accordingly, the cyber domain introduces unique challenges. For instance, scenarios must be presented that are unique and somewhat realistic, yet offer equivalent outcome measures of performance. Process measures must be identified and implemented that allow data to be collected in a non-obtrusive manner such that measurement does not interfere with participants exercising the skills and knowledge being measured. Furthermore, outcome and process measures must be identified that are generalizable to and predictable of performance within operational settings. By

beginning to address these issues, the proposed project advances the domain of cyber science through development of unique experimental methodologies, while providing a deeper understanding of situation awareness within the cyber domain. Furthermore, the current study offered an opportunity to collect data regarding secondary research questions concerning the effectiveness of cyber operations. Cyber security is a major challenge for DOE and other government agencies and there has been little scientific study of the human dimension of cyber operations.

Through the current study, data was collected that addressed group processes, the relationship between certain cognitive and personality attributes and the behavior and performance of cyber defenders, and the use of narrative in constructing stories to understand, explain and remember events in the cyber domain.

2 Methods

2.1 Participants

Thirteen employees from Sandia National Laboratories volunteered to participate in the experiment. All participants met the following requirements: (1). be 18 years or older, (2). have a background in computer science, (3). have an interest in cyber security/cyber incident response, (4). have not participated in any prior Tracer FIRE events and (5). be available on the designated dates for five full days of training and three full days to participate in the Tracer FIRE evaluation exercise.

2.2 Materials

The suite of network analysis tools used in the experiment included Encase Enterprise, Wireshark, IDA Pro, Volatility, Hex Workshop and PDF Dissector. Teams were additionally provided IRC chat as a means for intra-team communication and Plotweaver as an aide in creating a record of events.

2.3 Procedure

Participants were first asked to fill out a consent form and then complete a prescreening questionnaire. Next, the participants were asked to fill out a demographic questionnaire and a detailed questionnaire assessing general computer security and cyber incident response skills. This information was later used to assign individuals to the two training conditions and subsequently to place the participants into teams for the Tracer FIRE exercise. The objective was to assure that the three teams competing in the exercise were relatively balanced with respect to the knowledge and experience of team members.

Training. Participants were assigned to either the Tool-Based (7 participants) or the Narrative-Based (6 participants) training conditions. The two training groups received 3 days of training appropriate for their condition. The two training groups were then

combined for 2 additional days of training which addressed details concerning the use of the selected tools. This training was not as extensive as that provided in the Tool-Based training and emphasized the knowledge participants would need to solve the challenges in the Tracer FIRE exercise.

Tool-Based Training. Participants assigned to the Tool-Based training condition received 3 days of training focused on the functions incorporated into the tools and the mechanics of using the tools. This training involved relatively little information concerning adversary tactics and techniques and was comparable to training commonly provided by software vendors and included canned examples showing how the tools work, with relatively little emphasis on the application of the tools to real-world problems.

Narrative-Based Training. Participants assigned to the Narrative-Based training condition received 3 days of training emphasizing the theory of adversary tactics, application of tools and a detailed understanding of the role as a cyber incident responder. This training involved little consideration of the functionality of tools used for conducting network analysis. The training was structured in a manner that sought to help students comprehend the complex ideas and information in a form that was personal and formed relationships between their prior knowledge and personal experiences.

Tracer FIRE exercise. Following the 5 days of training, the participants were placed in one of three teams for the Tracer FIRE exercise. The Tool-Based team comprised of four participants whom had all received Tool-Based training. The Narrative-Based team comprised of four participants whom had all received the Narrative-Based training. The third group, the Combined team, composed of five participants; three of whom had received the Tool-Based training and two of whom had received the Narrative-Based training.

Each team was asked to solve multiple challenges to receive points with the score for each team continuously displayed and teams encouraged to compete against each other. The challenges were built around a coordinated series of events involving the same multi-level attack upon a host network of each team. The challenges required the teams to use the software tools addressed during training to analyze network traffic. This provided the basis for their interpreting events and establishing overall situational awareness. Points were awarded on the basis of successfully answering challenge questions concerning specific aspects of the attack, as well as their ability to form an accurate picture of the overall pattern of events (i.e., situational awareness).

Secondary Measures. Subjects were asked to complete a personality assessment consisting of the Big Five Inventory (BFI) from the website www.similarminds.com. Participants were also asked to perform three cognitive tasks: syllogism, comprehension span and mental rotation. These tasks have been used in previous studies and address different cognitive aptitudes associated with adaptive thinking and decision

making. The object was to assess whether these same aptitudes correlated with performance for cyber defender tasks.

Syllogism. This task is a measure of reasoning. Participants were given a logical argument in which a proposition is inferred from a set of premises and were asked to indicate whether the proposition was true given the premises.

Comprehension span. This task is a measure of verbal comprehension and associated memory recall. The participants saw a sentence and had to indicate whether the sentence made sense or not. After a series of sentences, the participant was asked to recall the last work of every sentence in order.

Mental rotation. This task is a measure of visual-spatial ability and mental flexibility. The participant was presented with a series of 20 pairs of figures. The task was to indicate whether the two figures, one of which was often rotated a specific amount of degrees, corresponded to the same object. The number correct that were classified in 60 seconds was taken as a measure of mental rotation ability.

Finally, at the beginning of the Tracer FIRE exercise, participants were told that there was a story embedded within the upcoming series of challenges. Furthermore, it was their task to discover this story as they solved the various challenges. It was encouraged that teams pay attention to cues associated with the stories and take notes to help them later piece together these cues. Then, at the end of the exercise, teams were given 30 minutes to construct an illustration depicting their interpretation of events and the underlying story.

3 Results

3.1 Descriptive Statistics

Participants were assigned to teams in a manner that provided a relative balance in the skills and experience of the individual team members. With respect to the questionnaire assessing general computer security and cyber incident response skills, the sums of the test scores for each team were Tool-Based training (Team 1) = 354, Combined training (Team 2) = 374, and Narrative-Based training (Team 3) = 347.

3.2 Training Type and Team Differences

The Narrative-Based team received the most points (11,182) followed by the Tool-Based team (10,480) and Combination team (9,811), respectively. This was also reflected in the average number of points received by team members; members of the Narrative-Based team individually scored more points on average than members of the other two teams.

A general linear model ANOVA with two factors was conducted to determine if there was a “training type” or “team” effect on the number of points obtained by teams. There were two levels in the training type: Narrative-Based or Tool-Based

training. There were three levels in the team factor: Team 1 (Tool-Based), Team 2 (Combined), and Team 3 (Narrative-Based) training. Neither training type nor team factor was significant and there was no statistical difference between team scores (i.e., "success") based on the training type or team.

3.3 Team Narratives

The teams were asked to prepare an illustration describing the story underlying the various events encompassed in the Tracer FIRE exercise. Figure 1 shows the ground truth depicted by the Plotweaver tool. As can be seen, there were multiple actors who intersected one another at key point in time. This was a multi-layered scenario that unfolded over time and it was not expected that any of the teams would be able to fully deduce all of the relationships that occurred across time and space.

Figure 2 shows the illustration prepared by the Narrative-Based team. It is apparent that this team failed to deduce many of the relationships between actors and events. However, this team did recognize five separate plot lines that loosely corresponded to those depicted in the ground truth storyline. Likewise, they recognized seven of the thirteen points at which the plotlines intersected one another. These two measures are believed to be indicative of the team's overall situation awareness which, as discussed below, was superior to that of the other two teams.

Figure 3 shows the illustration by the Combined team. This team deduced plotlines that loosely corresponded to four of the five plotlines within the ground truth depiction. Likewise, this team recognized the sequential development of events across time. However, it is striking that this team did not recognize any of the points where the individual plotlines intersected with one another. In fact, in both their hand-drawn illustration and their verbal account, the Combined team presented a linear sequence of events that did not involve any interactions between events, or individual actors. This team pieced together a story involving four separate actors that, for the most part, operated independently, when, in fact, the actors operated in concert with one another and this was a key element to interpreting the overall sequence of events. While this team clearly grasped the temporal structure of events, as well as the importance of individual actors, they were unable to deduce the relationships between different actors that were evidenced through their interactions as the scenario unfolded.

Figure 4 shows the illustration produced by the Tool-Based team. The Tool-Based team produced an even more impoverished illustration than either the Narrative-Based or Combined teams. They recognized three of the five plotlines. Yet, they recognized none of the relationships between the separate plotlines and two of the three plotlines that they did recognize consisted of a single event. Furthermore, their depiction captured none of the relationships between events or the relationships between different actors. Each member of this team seemed to have deduced one or more elements of the story independently; however, as a team, they were unable to put these elements together and did not seem to recognize that there was a coordinated action being taken by the adversaries.

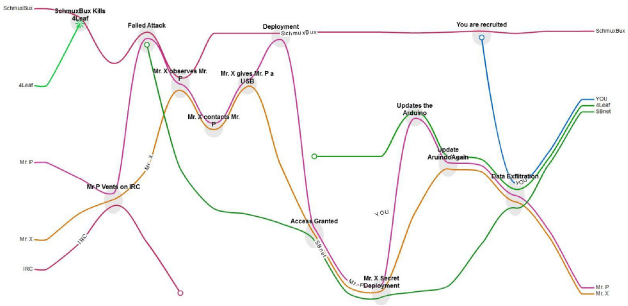


Fig. 1. Plotweaver Depiction of Ground Truth for Tracer FIRE Scenario

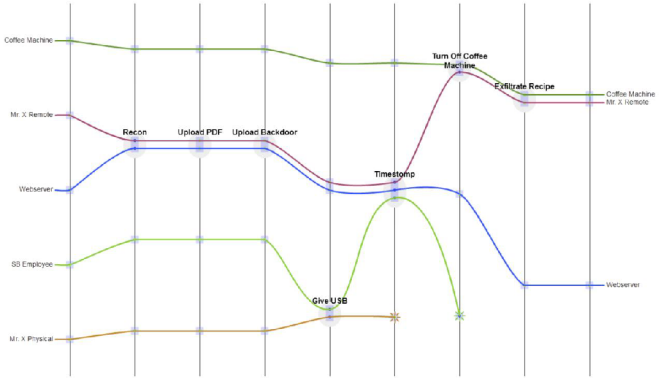


Fig. 2. Plotweaver Illustration Prepared by the Narrative-Based team

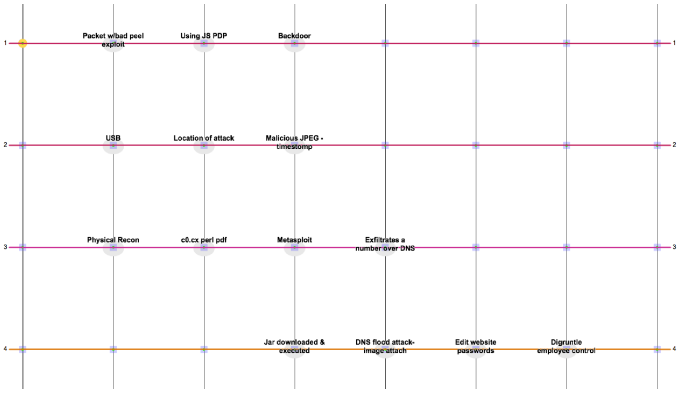


Fig. 3. Plotweaver Illustration Prepared by the Combined team

Interestingly, it was noted that all three teams deduced about the same number of story elements. During the Tracer FIRE exercise, there were specific challenges that, if successfully completed, teams learned a key element of the storyline. While the Narrative-Based team earned the most points in these challenges, there was not a huge

difference between the points earned by the Narrative-Based and the other two teams. This indicates that all three teams had many of the key story elements available to them but only the Narrative-Based team was able to put those story elements together in a way that corresponded to the actual relationships between events.

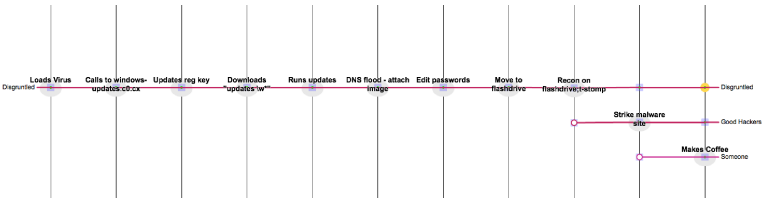


Fig. 4. Plotweaver Illustration Prepared by the Tool-Based team

3.4 Personality Factors

The BFI data was analyzed to determine if personality measures were associated with team success. Two subjects opted out of the personality assessment portion of the study. Therefore, only 11 participant's data was analyzed. A correlation matrix was calculated to determine if there was multicollinearity (Penney et al., 2011). The variables showed random scatter and no significant correlation.

A stepwise regression was conducted to determine if any of the variables were significant (alpha = 0.15, was set for selection in the stepwise regression). The final model included TimeTotal (total amount of time spent working on challenges), Inquisitiveness, and Emotional Stability. There were no departures from normality or outliers, and the residuals displayed constant error variance, with the error terms normally distributed. These data indicate that participants fell within the range that would be considered normal within the overall population and, therefore, results cannot be attributed to individual subjects with extreme scores on the Inquisitiveness or Emotional Stability personality dimensions. TimeTotal was not significant, but was included in the model as β1. Inquisitiveness and Emotional Stability were significant (R^2 = 58.87 and R^2-*adjusted* = 41.25, β0 = -7491, β1 =1.148e-12, *t*-value = 1.83, *p*-value < .1093). Inquisitiveness was marginally significant (β2 = 63, *t*-value = 2.02, *p*-value= .083) as was Emotional Stability (β3 = 88, *t*-value = 2.98, *p*-value= .021). Only Emotional Stability was included in the final model.

3.5 Cognitive Factors

The three cognitive tasks, Mental Rotation (MRScore), Comprehension Span (CompS) and Syllogism (Syllo), were analyzed to determine if they were associated with team success. Four subjects opted out of the cognitive task portion of the study. Therefore, only 9 participant's data was included.

The final model included CompS (R^2 = 47.25, R^2-*adjusted* = 39.71, β0 = 1459, β1 = 32 with a *t*-value =2.50, *p*-value <0.041). There were no departures from normality, no outliers, the residuals displayed constant error variance, and the error terms were

normally distributed. Thus, the results could not be attributed to individual subjects who exhibited extreme scores, as all subjects were within the range that would be considered normal for the population.

4 Conclusion

The results from this study provide insights concerning alternative methods for delivering training for cyber defenders, as well as a better understanding of factors contributing to team situation awareness and individual and team performance of cyber defenders. Most notably, this study highlights the importance of the narrative, or the capacity to interpret events and put them into the context of a story, to the effective use of software tools by cyber defenders. Furthermore, the study also illustrates the importance of individual characteristics to the ability of individuals to effectively work together within a cyber incident response team.

With only three teams, it was not possible to demonstrate a statistically significant difference in the performance of the teams receiving alternative modes of training, although the team receiving Narrative-Based training did earn more points than their counterparts. Likewise, on average, the members of the Narrative-Based team individually earned more points than their counterparts on the other teams. While not statistically significant, these results are in the expected direction and are consistent with detailed analysis of overall situation awareness exhibited by the three teams.

Assessments of personality and cognitive factors revealed two variables that were significantly correlated with individual performance during the cyber exercise. With respect to personality, those who exhibited higher scores on the Emotional Stability dimension performed better. Those scoring high on this dimension tend to be more secure and confident, whereas those scoring low exhibit a greater tendency to show unpleasant emotions such as anger, anxiety, depression and vulnerability. It should be noted that while the participants in the current study exhibited a range of scores on this dimension, their scores fell within the range considered normal for the overall population.

There are two important ramifications for the finding that individual performance correlated with Emotional Stability. First, during training, the Emotional Stability of individual students may be expected to affect both the benefit derived from the training experience, as well as the performance during training exercises, such as Tracer FIRE. Thus, it is proposed that mechanisms be employed that allow individual and team performance to be more closely monitored in real-time so that instructors may effectively intervene when students have become non-productive and are struggling. Likewise, in composing teams, it may be beneficial to combine individuals with varying experience and maturity to provide some degree of scaffolding for weaker team members who may become easily discouraged.

Second, and perhaps more importantly, within operational settings, it may be expected that personnel will exhibit varying levels of Emotional Stability and this will have an indirect, and perhaps direct, effect on their performance. This may be manifested in their capacity to effectively function within teams, as well as their capacity

to cope with ongoing stressors. It is uncertain what countermeasures may be most appropriate; however this represents an important consideration given the nature of the Cyber domain where technically qualified personnel are in high demand and many organizations find it difficult to retain their best talent.

A second individual factor that correlated significantly with performance was Comprehension Span. In this task, subjects were presented a series of sentences and after each sentence, they were required to indicate if the sentence made sense. Then, their memory span was tested by requested that they recall the last word in each sentence. To perform well, an individual must have both proficient at interpreting verbal content and possess good short-term memory. Previous studies have shown that individuals who perform well on this measure also perform well in tasks requiring adaptive decision making. Here, adaptive decision making is defined as the capacity to recognize that a strategy is ineffective and thus, there is need to either alter an existing strategy or abandon an existing strategy for an alternative strategy (Abbott et al., 2011). It is proposed that the challenges presented through the Tracer FIRE exercise place similar demands for adaptive decision making upon the participants and that Comprehension Span represents a fundamental cognitive attribute underlying effective performance.

References

1. Abbott, R., Haass, M., Trumbo, M., Stevens-Adams, S., Hendrickson, S., Forsythe, C.: Robust Automated Knowledge Capture, SAND 2011-8448, Sandia National Laboratories (October 2011)
2. Penney, L.M., David, E., Witt, L.A.: A review of personality and performance: Identifying boundaries, contingencies, and future research directions. Human Resource Management Review 21, 297–310 (2011)
3. Tadda, G.P.: Measuring the Performance of Cyber Situational Awareness Systems. In: Proceedings of the 11th International Conference on Information Fusion, Cologne GE, June 30-July 3 (2008)

Instrumenting a Perceptual Training Environment to Support Dynamic Tailoring

Robert E. Wray, Jeremiah T. Folsom-Kovarik, and Angela Woods

Soar Technology, Inc., 3600 Green Court Suite 600,
Ann Arbor, Michigan, USA, 48015
{wray,jeremiah.folsom-kovarik,angela.woods}@soartech.com

Abstract. Simulation-based practice environments would be more valuable for learning if they supported adaptive, targeted responses to students as they proceed thru the experiences afforded by the environment. However, many adaptation strategies require a richer interpretation of the student's actions and attitudes than is available thru the typical simulation interface. Further, creating extended interfaces for a single application solely to support adaptation is often cost-prohibitive. In response, we are developing "learner instrumentation middleware" that seeks to provide a generalized representation of learner state via reusable algorithms, design patterns, and software.

Keywords: Perceptual learning; adaptive training; learner modeling.

1 Introduction

Many of today's computer-based learning environments offer simulacrums of the performance environment. These practice environments enable a learner to practice skills and to demonstrate knowledge of concepts that are the subject of training, offering support for more sustained and thus potentially deeper and more complete learning [1-3]. Although theoretical debate continues regarding how to best structure practice experiences, an emerging consensus agrees that dynamic adaptation of practice to enable targeted, individualized experience is important for effective computer-based training [4].

We are taking an applied perspective to practice environments, focusing on delivering effective and adaptive instruction thru whatever means appears apt for the domain. Toward this end, we are developing general learner instrumentation and tailoring capabilities that enable practice environments to adapt to the learner both extrinsically (outside of the domain experience of the simulation) and intrinsically (within the simulated experience). These capabilities also are designed to support both learner cognitive and affective states. The resulting Dynamic Tailoring System [5, 6] is designed to integrate instructional methods and best practices as they are identified and validated and also serves as a testbed for researching such adaptation strategies.

The Dynamic Tailoring System (DTS) has been demonstrated in multiple practice domains. Each domain imposes specific requirements. Although many requirements

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 100–109, 2013.

are shared across domains, some are unique. A core engineering challenge is to define and implement a capability that is sufficiently functional and flexible to support the common and the unique requirements of a particular application. In this paper, we explore this tension, focusing on the general challenge of learner instrumentation: processing and packaging inputs from a learner and simulation to enable the classification/recognition of learner states. These states then help the system identify the best interventions for the individual student at that moment. To illustrate with a concrete example, we introduce a perceptual training application that demands more powerful instrumentation than earlier applications because the human practice task is primarily one of observation, in which explicit action, which is easier to recognize and assess, is relatively infrequent.

In response to these limitations, we are developing "learner instrumentation middleware" that transforms, supplements, and fuses simulation and learner data into a succinct representation of learner context and state. We describe requirements of the learner-instrumentation capability such as extensibility and reusability. We then describe how we are developing and applying this middleware in the context of the perceptual training application.

2 Enabling Dynamic Tailoring

We have been developing the Dynamic Tailoring System as a general-purpose software architecture for dynamic tailoring; that is, pedagogical experience manipulation during practice [1, 6]. The system is specifically designed to support many types of simulation-based training systems and domains. We have implemented a functional implementation of this system and have demonstrated it in multiple domains. Figure 1 illustrates the current architecture. There are three core functional components (boxes) and four primary representational components (database icons). Here, we briefly outline the overall design of the system in order to motivate and to provide context for the learner instrumentation challenge below.

Monitor. The Monitor observes learner actions, interprets those actions in the context of the learning situation (via a domain/expert model), assesses the learner's behavior in terms of active learning objectives, and then classifies the observed behavior using a behavior ontology. As we outline further below, the Monitor is supported by several translation layers ("learner instrumentation middleware") that decouple the details of simulation environments and learner sensing from the representations used for interpretation.

Pedagogical Manager. The Pedagogical Manager maintains an estimate of proficiency for each learning objective, decides between extrinsic mediation (such as an ITS dialog) and the intrinsic tailoring, and chooses alternative instructional strategies. For example, a "scaffold" tailoring strategy can be used when a learner has demonstrated high levels of competence but is transitioning to more complex challenges or new learning objectives within the domain [7]. The Pedagogical Manager also mediates choices between affective and domain-content tailoring strategies.

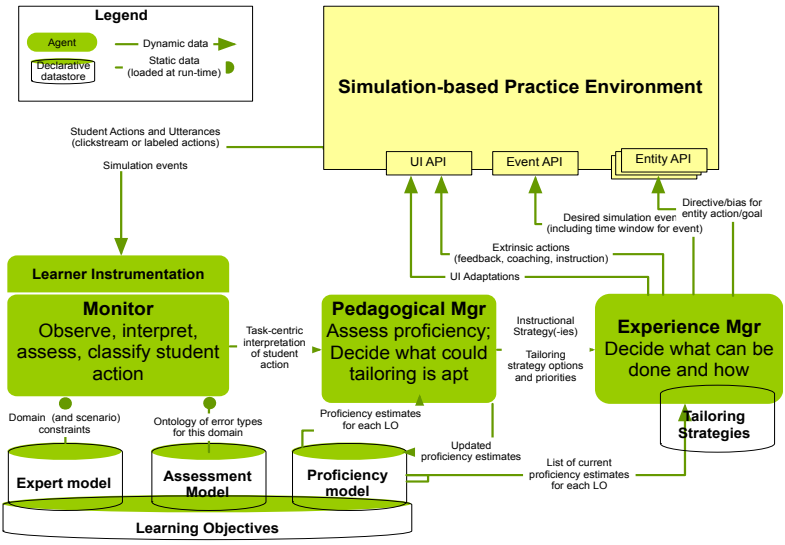


Fig. 1. Architectural composition of the Dynamic Tailoring System

Experience Manager. The Experience Manager chooses and instantiates tailoring strategies based on general recommendations from the Pedagogical Manager. For example, the Pedagogical Manager may recommend a tailoring strategy that is intrinsic, meant to challenge, and focuses on several enumerated learning objectives. The Experience Manager then evaluates tailoring strategy options to determine which strategies can be used to satisfy the request.

3 Instrumentation Requirements for Perceptual Training

The previous section outlined a high-level software architecture to support dynamic tailoring across a range of training applications. In this section, we wish to examine architectural requirements more substantively, focusing on a specific training application to highlight requirements. The training application includes a practice environment in which US Marines observe a village from a Virtual Observation Post (VOP).

The VOP is inspired and informed by successful "live" training programs [8, 9]. In this training, Marines learn to construct a general "baseline" of understanding from sustained attention to the activities in a "village" (populated by human role players). Marines exchange observations with one another and practice the application of observational skills introduced in a classroom. The resulting sensemaking skill covers a broad range of perceptual skills, from low-level signals (recognizing the proxemics and kinesics or "body language" of individual villagers), to recognizing and categorizing quotidian and unusual events, to developing an abstract mental representation of the patterns of life within the village.

For the VOP, the implementation of adaptive tailoring strategies is comparatively straightforward [10]. The learner is positioned in a Virtual Observation Post 1000m or

more from the observed village. As a consequence, learner action requires less interaction and coordination with the simulation than in a domain in which a learner would be directly interacting with a business partner, an accident victim, or aircraft in a virtual battlespace. Learner actions include focusing "optics" (e.g., binoculars) on specific locations in the scene, reporting observations and events, and suggesting interpretations of events for others to consider and discuss. Open-ended speech recognition across a team of learners is a challenging technical problem, but, from the point-of-view of the tailoring system, these inputs are pre-processed to as labeled text strings.

Although learner actions are comparatively simpler, it is also relatively more challenging to develop and maintain an understanding of learning state in this domain. A learner may spend many minutes just scanning a scene with binoculars. During this time, numerous observations may be made (or missed) by the learner without an explicit utterance or formal report. An understanding of learner state is necessary for deciding what tailoring actions are relevant at a particular time for a particular learner/team. Without a good understanding of the learning state, appropriate and timely instructional tailoring is not possible. Worse, inapt tailoring may also increase learner frustration and negatively impact learning.

The requirements for tailoring in turn impose additional constraints on the practice environment. A practice environment without instructional supports (such as tailoring) needs only to allow a learner to take action in the environment (reporting an event to other team members, choosing different sensing optics, firing a weapon, etc.). Richer instrumentation of the practice environment is necessary to enable interpretation of a learner's actions and maintenance of a dynamic and reasonably accurate model of the learner.

Instrumentation is difficult because simulation affordances for learner observation are typically weak. Most simulations provide minimal descriptions of learner activity and without directly providing learner/task context; e.g., they may indicate that a

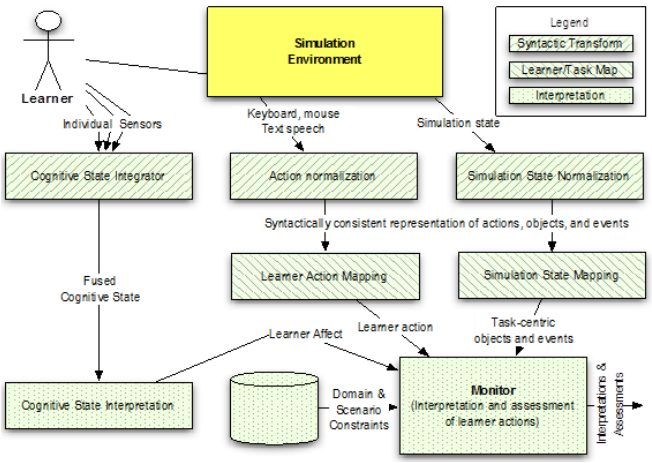


Fig. 2. Conceptual design of the learner instrumentation middleware

learner took some action at some time, but cannot indicate the appropriateness of that action. A related limitation is the available "channels" of observation. Most simulations provide only keystroke and mouse inputs, which are much more limited than the kinds of signals and information a human tutor might get from interaction with a learner.

4 Learner Instrumentation Middleware

In order to support adaptive training, the simulation environment requires additional methods of learner instrumentation in addition to just a practice environment. However, these requirements have the potential to add significant cost to the training system. The approach we have been investigating and exploring is "learner instrumentation middleware." This section outlines the design of this middleware, current progress toward the goal, and some of the tradeoffs in pursuing middleware versus application-specific solutions.

Fig. 1 illustrates the conceptual design of the learner instrumentation middleware. This software seeks to transform, supplement, and fuse simulation and learner data into a succinct representation of learner context and state with a general (not domain specific) set of functions and processes. The potential advantage of the middleware is that it can collect and transform individual sensor and input streams from the learner and simulation into representations that are largely independent of these sources. This approach allows the interpretation and adaptation algorithms used in the remainder of the Dynamic Tailoring System to be independent of the specific simulation environment and sensor suite.[1] There are three distinct layers to the middleware: 1) syntactic normalization, 2) learner/task mapping, and then 3) interpretation. Each of these layers is sketched individually below.

4.1 Syntactic Normalization

This layer converts the specific representations used by the simulation to a general representation used within the Dynamic Tailoring System. In the examples in this paper, we use a predicate representation of the normalized syntax for simplicity/clarity. This layer is largely custom-built for each simulation environment. However, this layer is not application specific, meaning that components in this layer can be reused for different training applications using a common simulation environment.

The Cognitive State Integrator (CSI) is a more sophisticated component than the other two in this layer. The goal of the CSI is to provide a consistent representation of estimated cognitive states regardless of the sensor(s) used to measure indices of these states. At this level of the middleware, indices measured from different sensors are fused and leveled, providing an estimate of a particular cognitive-state dimension (such as arousal or attention) at the current moment in time.

[1] An "inverse transform" is required for translation of adaptive interventions into simulation-specific functions.

4.2 Learner/Task Mapping

The role of this layer is to translate or map the outputs from the syntactic normalization layer to a representation that is focused on the learning task rather than simulation events. The transformation at this layer is in some respects analogous to an affine transformation, in that the transformation enables the consumer of the transformed data to be simpler. The mapping process simplifies and unifies the range of inputs the interpretation layer must consume, allowing interpretation to be domain neutral. Ideally, it will also be possible for the mapping layer to be reusable across domains. However, we have not yet developed general, reusable algorithms for this layer and it thus currently requires custom programming for each new application. We are currently investigating a scenario representation with a formal (ontological) representation within this layer to reduce and simplify the custom development requirements.

Several examples of the kinds of mapping provided in this layer are summarized in Fig. 3. Imagine a situation where a learner is tasked to track and report on the actions of an individual moving thru a small village. At some point, this actor enters a marketplace. The learner, using virtual binoculars with high magnification, notices that the high-interest individual is visibly angry and reports that over a simulated radio. The syntactic layer, as above, converts data from the different components of the simulation – an optics simulation, a simulation environment (e.g., VBS2), speech-to-test components – and converts them into to a predicate representation similar to that pictured in Fig. 3.

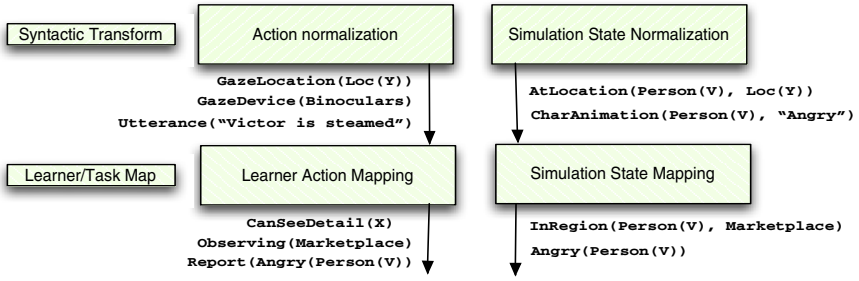


Fig. 3. An example of inputs and outputs for the Learner/Task Mapping layer

This translation decouples the syntactic details of the different system components for interpretation but representations remain tied to their original frames. For example, the syntactic layer will provide data about the location of particular objects, but not how those objects relate to the learner and the learning task. The mapping layer provides this translation. As suggested by the figure, the combination of the use of binoculars and focus on the marketplace allows the system to be able to infer that the learner `CanSeeDetail()` of objects in the market, such as the angry expressions of the high-interest individual. As we discuss further below, these mappings to the learning context make it much simpler for the interpretation layer to reason about the learning situation and assess a learner's action(s).

4.3 Interpretation and Assessment

Monitor. The syntactic and semantic transform layers feed the interpretation and assessment layer. As outlined above, the "Monitor" evaluates the current learning state, as represented by the outputs of the semantic transform layer against a collection of user-defined constraints. The Monitor is implemented using the Soar architecture as an agent architecture [11] and takes advantage of a highly efficient pattern matcher to evaluate the constraints against the learner-oriented description of the situation provided by the previous layers. These constraints were originally inspired by constraint-based expert modeling [12] but have been extended and customized for this function in the Dynamic Tailoring System. One specific example of a customization is a codification of distinct domain, scenario, and practice constraints [13], which enables (as one example) the monitor to assess the same learner action differently based on the specific goals of a practice exercise.

Figure 4 illustrates how the mapping and interpretation functions of the middleware components enable improved generality, ease of authoring, reusability, and transparency for the Monitor. Continuing the example from above, the Monitor's rules can leverage the general predicates `Angry()` and `InRegion()` to test for classes of events that should be reported, rather than needing to include simulator-specific tests for specific character grid locations and animations. The middleware lets the Monitor query simulation-specific inputs such as simulation state or physiological sensors and easily interpret the learner's behavior in order to determine not just whether the learner reacted correctly, but what underlying reasons might have caused any incorrect outcomes. The outputs of the Monitor (*shaded boxes*), which drive pedagogical decisions in the DTS, can be specified more abstractly allowing instructors to understand and control the system's behavior. Finally, the Monitor rules can be reused when new scenarios or new sensor input sources are added.

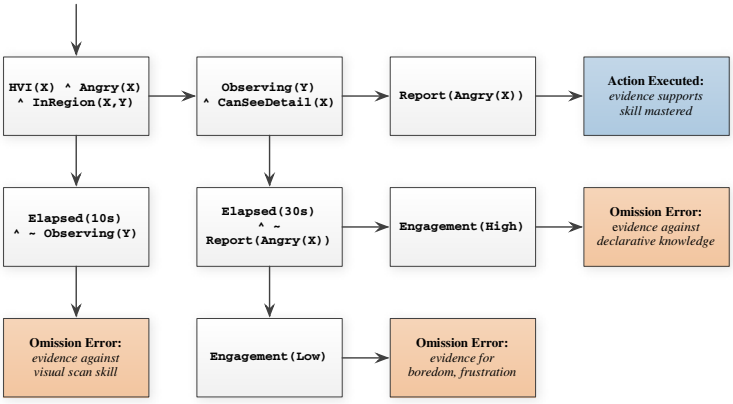


Fig. 4. Generalized predicates within the Monitor are able to describe a wide range of learner behaviors independently of simulation-specific details such as line-of-sight calculation and individual physiological sensor inputs

A significant benefit of the learner middleware approach is that the Monitor itself can be reused from one application to another without significant recoding. We have used the Monitor for applications as diverse as cross-cultural conversation, military decision-making and medical triage. A primary impetus for formalization of the middleware, as described in this paper, is our recognition that the constraint-based representation and pattern-matcher is proving powerful for many different applications.

Cognitive State Interpretation. The cognitive state interpretation function supports the Monitor but also provides direct measures of learner cognitive state and/or affect to other components of the DTS. Fig. 5 illustrates the cognitive state interpretation function. In this example, the dimension of interest is arousal/attention. The syntactic layer fuses sensor inputs and places individual observations on a normalized attention axis at a particular time, as outlined previously. The interpretation layer then compares the observations to a bounding "envelope" that defines the minimum and maximum desired levels for the dimension at a particular time.

The envelope provides a simple to use, actionable interpretation of cognitive state for other DTS components to use. An individual observation (or a prediction based on the trend/derivative) can help the DTS understand the relative priority and urgency of affective interventions. In the example in the figure, the learner's falling attention and the proximity of the current attention level to the lower bound of the envelope may lead the DTS to prioritize an attention-oriented tailoring strategy over a conceptual one. Similarly, the Pedagogical Manager might recommend an extrinsic intervention rather than in intrinsic adaptation in this situation because attention is sufficiently low that there is likely to be little interference with the learner's sense of presence in the practice experience.

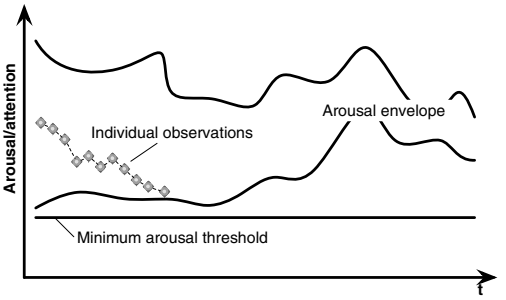


Fig. 5. Illustration of cognitive state interpretation

These envelopes today must be constructed by hand and adjustments made for different levels of proficiency (i.e., the envelope for the same task may be different for learners with different estimated levels of competency in the task). However, we are interested in methods that would allow us to construct them automatically and to compose envelope segments to accommodate learner actions and branching events within a scenario.

5 Conclusions

This paper has presented the conceptual design of general-purpose abstraction and translation layers to make it easier to obtain richer information from a practice environment than is typically afforded by a simulation environment. Although we noted several areas where the current implementations of this middleware are not yet fully developed or limited in their generality, the development thus far is providing benefit. We see two primary advantages to this learner-instrumentation middleware. First, it lowers the cost of integrating adaptive tailoring into a practice environment. Cost is reduced by supporting faster and simpler integration with simulation environments and by enabling reuse of the primary DTS components (Monitor, Pedagogical Manager, and Experience Manager) across applications. Second, it enables the integration of additional learner information streams, such as cognitive and affective state. The hypothesis is that these additional sources of information will enable more accurate diagnosis of the learner's needs and progress, extend the range of adaptation, and, ultimately, improve the efficiency of training.

Acknowledgements. This work is supported in part by the Office of Naval Research. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. This work was supported, in part, by the Office of Naval Research project N00014-11-C-0193, Perceptual Training Systems and Tools (PercepTS).

References

1. Lane, H.C., Johnson, W.L.: Intelligent Tutoring and Pedagogical Experience Manipulation in Virtual Learning Environments. In: Cohn, J., Nicholson, D., Schmorrow, D. (eds.) The PSI Handbook of Virtual Environments for Training and Education, vol. 3, Praeger Security International, Westport (2008)
2. Schatz, S., Oakes, C., Folsom-Kovarik, J.T., Dolletski-Lazar, R.: ITS + SBT: A Review of Operational Situated Tutors. Military Psychology, special issue on current trends in adaptive training for military application (2012)
3. Schatz, S., Bowers, C.A., Nicholson, D.: Advanced situated tutors: Design, philosophy, and a review of existing systems. In: 53rd Annual Conference of the Human Factors and Ergonomics Society, pp. 1944–1948. Human Factors and Ergonomics Society, Santa (2009)
4. Tobias, S., Duffy, T.M. (eds.): Constructivist Instruction: Success or Failure? Routledge, Taylor and Francis, New York (2009)
5. Wray, R.E.: Tailoring Culturally-Situated Simulation for Perceptual Training. In: Duffy, V. (ed.) Proceedings of the 2012 Applied Human Factors and Ergonomics Conference, 2nd International Conference on Cross-Cultural Decision Making: Focus 2012. CRC Press, Taylor and Francis, Boca Raton, FL (2012)

6. Wray, R.E., Lane, H.C., Stensrud, B., Core, M., Hamel, L., Forbell, E.: Pedagogical experience manipulation for cultural learning. In: Workshop on Culturally-Aware Tutoring Systems at the AI in Education Conference, Brighton, England (2009)
7. Lane, H.C., Wray, R.E.: Individualized Cultural and Social Skills Learning with Virtual Humans. In: Durlach, P.J., Lesgold, A.M. (eds.) Adaptive Technologies for Training and Education, pp. 204–221. Cambridge University Press, New York (2012)
8. Schatz, S., Reitz, E., Nicholson, D., Fautua, D.: Expanding Combat Hunter: The science and metrics of Border Hunter. In: Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL (2010)
9. Gideons, C.D., Padilla, F.M., Lethin, C.R.: Combat Hunter: The training continues. Marine Corps Gazette, pp. 79–84 (2008)
10. Schatz, S., Wray, R., Folsom-Kovarik, J.T., Nicholson, D.: Adaptive Perceptual Training in a Virtual Environment. In: Human Factors and Ergonomic Systems (HFES 2012), Boston (2012)
11. Wray, R.E., Jones, R.M.: An Introduction to Soar as an Agent Architecture. In: Sun, R. (ed.) Cognition and Multi-agent Interaction: From Cognitive Modeling to Social Simulation, pp. 53–78. Cambridge University Press, Cambridge (2005)
12. Mitrovic, A., Ohlsson, S.: Evaluation of a constraint-based tutor for a database language. International Journal of Artificial Intelligence in Education 10, 238–250 (1999)
13. Wray, R.E., Woods, A., Priest, H.: Applying Gaming Principles to Support Evidence-based Instructional Design. In: 2012 Interservice/Industry Training, Simulation, and Education Conference, Orlando (2012)

Part II

Team Cognition

Improving Tool Support for Software Reverse Engineering in a Security Context

Brendan Cleary[1], Christoph Treude[1], Fernando Figueira Filho[1], Margaret-Anne Storey[1], and Martin Salois[2]

[1] Dept. of Computer Science, University of Victoria
Victoria, BC, Canada
bcleary@uvic.ca
[2] Defence Research and Development Canada – Valcartier
Quebec, QC, Canada
martin.salois@drdc-rddc.gc.ca

Abstract. Illegal cyberspace activities are increasing rapidly and many software engineers are using reverse engineering methods to respond to attacks. The security-sensitive nature of these tasks, such as the understanding of malware or the decryption of encrypted content, brings unique challenges to reverse engineering: work has to be done offline, files can rarely be shared, time pressure is immense, and there is a lack of tool and process support for capturing and sharing the knowledge obtained while trying to understand assembly code. To help us gain an understanding of this reverse engineering work, we conducted an exploratory study at a government research and development organization to explore their work processes, tools, and artifacts [1]. We have been using these findings to improve visualization and collaboration features in assembly reverse engineering tools. In this talk, we will present a review of the findings from our study, and present prototypes we have developed to improve capturing and sharing knowledge while analyzing security concerns.

Keywords: malware, reverse engineering, empirical study.

1 Introduction

In his 1987 article [3], Cohen coined the term "computer virus" to describe self-reproducing programs designed to infect other computer programs. At that time, computer viruses were created for experimentation purposes or merely for fun, therefore causing little damage to real world systems [6].

Today's landscape shows us a different scenario. Computers are widely used in criminal activities such as bank fraud, identity theft, and corporate theft. According to a recent Symantec report [10], 2011 saw more than 187 million identities exposed in data breaches caused by hacking, and 93% more vulnerabilities related to mobile platforms—up to 315 in 2011 from 163 in 2010.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 113–122, 2013.

Illegal activities in cyberspace affect national security and threaten citizen's rights and privacy, thus having significant political, economic, and social implications [1]. Organized cyber groups typically communicate using cryptographic protocols and store information using encrypted files or systems. As a countermeasure against cybercrime, government institutions and business organizations have been using reverse engineering methods to analyze malicious code and break into password protected file systems.

This paper summarizes the findings of a first-of-its-kind field study we conducted with security engineers working in a government research and development organization [11] and looks at how we are incorporating these findings into tools designed to assist security engineers performing exploitability analysis [2].

Our study shows that security engineers have a unique work environment and experience significant challenges with urgency, documentation, and a limited ability to share information. Overall, security engineers have special needs in terms of time sensitivity, coordination, communication, and documentation.

2 Field Study Overflow

Our field study was conducted as an exploratory qualitative study. We conducted seven semi-structured interviews with engineers at a government research and development organization tasked with understanding targeted malware. For the remainder of this paper, we use P1 to P7 to refer to the participants of our study. A full description of the methodology can be found in [11]

To gain a comprehensive understanding of software reverse engineering in a government security context, our research questions focus on processes, tools, and artifacts:

1. What processes are part of reverse engineering in a security context?
2. What tools are being used?
3. What artifacts are being created and shared?

3 Summary of Findings

In this section, we present a summary of our findings from [11], subdivided for each research question posed in Section 2.

3.1 Processes

Based on the interview data, we identified five processes that are part of reverse engineering in a security context.

Analyzing. Analyzing assembly code is at the heart of most reverse engineering projects. Typical projects include the detection of malware, such as trojan horses, or the decryption of encrypted file systems. Assembly code is more difficult to understand than source code written in high-level programming languages

because the code is less structured, often lacks meaningful symbols or data definitions, and allows for tricks that can mislead reverse engineers in their analysis efforts. Following the flow of data is challenging: *"Understanding the data flow is a big part of understanding a program."* $_{P4}$

Documenting. Documenting reverse engineering has several purposes. Some documentation is done to provide cognitive support for the reverse engineers at the time of the analysis, some documentation is meant to capture the reverse engineers' own understanding of the code, and other documentation is meant to be shared either with team members or outside stakeholders. While it is already difficult to document source code written in high-level programming languages, it is even more difficult when dealing with assembly code. During the exploration of the assembly code, most reverse engineers document just enough information to be able to resume a task and do not document the paths that were explored without success.

Transferring Knowledge. Transferring knowledge is a challenge in reverse engineering. Documentation alone is often not enough to understand the work that has been completed by somebody else: *"[I would] look at a version with comments, but I'd still need to jump through to understand."* $_{P7}$ In the current setting, information is usually passed on verbally or via email and chat. These mechanisms do not scale beyond groups of about five reverse engineers. To solve some of these issues, the idea of a workflow would be useful: *"Right now it's being done like a craft, and we'd like to have some kind of assembly line"* $_{P4}$. However, workflows are not consistent for all cases, and most workflow support tools are too constraining. In addition, documentation conventions and information sharing standards could improve the reverse engineering process: *"Respecting conventions [would make it] easier to pass from one project to another."* $_{P2}$

Articulating Work. Articulating work consists of all the items needed to coordinate a particular task, including scheduling sub-tasks, recovering from errors, and assembling resources [4]. In reverse engineering, where tangible results are only produced when a path of exploration is successful, constantly re-doing work is a problem. Work was usually divided based on different pieces of hardware, different vulnerabilities, different functions, or different files. Relating information from the analysis of different pieces of the problem was very difficult.

Reporting. When external stakeholders are involved, the final step in a project is reporting the results of the reverse engineering activities. In some cases, reporting includes a great deal of articulation work, especially when artifacts can be co-opted as reports: *"Instead of writing a report we shared a Word document."* $_{P6}$

3.2 Tools

Tools used by the participants in our study can be classified as disassemblers, office productivity and visualization tools, and communication and coordination tools.

Disassemblers. Most of the reverse engineering work is performed using IDA Pro[1]. IDA Pro is a commercial product that performs automatic code analysis and offers interactive functionality to support the understanding of disassembly. Reverse engineers typically start with an automatically generated disassembly listing, then rename and annotate sections in the listing until they understand the code. Debuggers are rarely used for malware in the early stages of analysis since portions of the code required for execution are often missing or because of the need to first remove anti-debugging tricks used by the malware. As one of our interviewees described it, the main analysis tool used by reverse engineers in the security context is *"brain power"*$_{P6}$.

Office Productivity and Visualization Tools. Most of the documentation is written using Microsoft Word, Excel, or OneNote. UML sequence diagrams are usually drawn to represent control flow understanding. However, the reverse engineers had *"trouble finding good tools that draw graphs and make it easy to navigate and export graphs"*$_{P1}$. Paper was also used, primarily for workflow support, small graphs, and articulation work.

Communication and Coordination Tools. For communication, only basic tools, such as e-mail and chat, were used. Our interviewees work in a co-located setting that allows face-to-face communication, but data sharing is complicated by the nature of the classified work. Interviewees coordinated work using tools such as wikis, bug trackers, and shared documents.

3.3 Artifacts

Artifacts created during the reverse engineering process in our setting consist of annotations, artifacts created for cognitive support, and reports.

Annotations. IDA Pro supports two notions of annotations: repeatable and non-repeatable. A repeatable annotation will appear attached to the current item as well as other items referencing it. Non-repeatable annotations only appear attached to the current item[2]. In addition, pre-comments and post-comments can be attached to lines and functions. All annotations also show up in the IDA Pro dependency graph.

The reverse engineers used annotations for several reasons: to keep track of variables, to rename functions, to document jumps, and to record where a particular piece of code was reading from or writing to. However, one of the challenges is that annotations are always incomplete: *"When you document stuff you tend to skip stuff that's obvious at the time."*$_{P6}$

Cognitive Support Artifacts. Depending on the use case, different documents are created by the reverse engineers to aid their cognition. These include:

[1] http://www.hex-rays.com/idapro
[2] http://www.hex-rays.com/idapro/idadoc/480.shtml

memory maps, Excel or Word tables showing register usage and boot processes, data flow diagrams, sequence diagrams, and scripts. A common scenario is when an engineer needs to keep track of different paths that are being explored in order to understand a particular piece of code. One of our interviewees used Microsoft OneNote to do that: *"I also used OneNote in other projects to keep track of paths that way. The last line in the OneNote document was the last path [that I had] explored."*$_{P6}$

Reports. Companies focused on malware, such as Symantec, frequently create reports that provide an overview of how a particular piece of malware works. Such reports rarely include enough detail to understand the inner workings of the malicious program, mostly because security companies do not want to reveal their insights to malware writers. In contrast, reports produced in our study setting had more technical content, and often included assembly code for functions as well as detailed descriptions of all input and output parameters.

4 Discussion of Challenges

Each work process described in the last section involved a different set of tools. These tools, in turn, were used to produce artifacts in distinct, non-interoperable formats. Therefore, moving from one process to another required a lot of manual work. By moving from the analysis to the documentation, engineers produce artifacts that would help them resume their own tasks, as well as transfer their knowledge to other team members. For example, reverse engineers have tried using wiki-based systems for sharing mixed content (e.g., details on how particular hardware works, including pieces of code). However, wikis have shortcomings when navigating code and related artifacts: *"Wikis are very document like, not ideal for documenting code – some kind of graph tool would have been better."*$_{P1}$. Overall, even when knowledge sharing was encouraged, reverse engineers faced a lack of proper tools to pass information along to others: *"There's also stuff that we don't know how to document."*$_{P1}$. Navigation is particularly a challenge when dealing with different documents such as the cognitive support artifacts mentioned above. A map of all documents and their connections usually only exists in the reverse engineer's head.

To articulate their work and break problems into pieces, engineers often followed a divide-and-conquer strategy: *"We go after different pieces. The problem is how to share information then... different people have different processes."*$_{P2}$. This poses an interesting phenomenon: there is no general process in the work of security reverse engineers. The following factors would influence this phenomenon:

Task Complexity. Tasks, such as blocking malware and breaking into secure devices, often include unsolved problems, thus requiring the use of different approaches, tools, and skills.

Security. The security context further obstructs the reverse engineers' work. Classified information cannot be easily shared, and for classified tasks, the reverse engineers are only allowed to work on classified, often un-networked, equipment. Often, information cannot be transported since it could belong to different projects, security classifications, or machines. Even for unclassified contexts, such as malware, the nature of the code prohibits easy sharing to prevent further infection. This also means that a lot of the work has to be completed offline, and access to web resources is very limited. Most of the reverse engineers in our study worked by themselves, often for security reasons: *"I'm the only one allowed to look at it [...] You don't want others to be infected [with malware]"* $_{P2}$.

Time Constraints. The amount of time pressure depends on the scenario. Some projects have the goal of understanding everything about a particular piece of software and are usually completed without time pressure. In other scenarios, only a couple of weeks are allocated for a particular project in order to provide a fast response to a potentially harmful threat. In the latter case, the reverse engineers have to prioritize what they are working on. In the example of malware: *"[We have] four goals when dealing with malware: detect, block, remove, [and] understand everything. Usually [the process] stops after the third step."* $_{P7}$ The amount of documentation produced depends on the extent of the time pressure. Long-term projects without time pressure yield more documentation, whereas for short-term projects, there is often not enough time to document thoroughly: *"If you put too much documentation, you won't have enough time to finish."* $_{P2}$

Tool Constraints. A graph is often the best way to capture a certain aspect of a reverse engineering problem, but it is difficult to deal with different types of diagrams. One of our interviewees told us that he sometimes spends up to 100 hours creating a single diagram. Also, the graphs produced are usually not linked to the disassembly, thus losing traceability. There is a shortage of tools that span different aspects of reverse engineering, such as hardware specifications and assembly code. The reverse engineering is also limited by memory since tools rarely scale beyond executables larger than a few megabytes.

5 Application of Findings

To demonstrate how these insights into the work practices of security engineers can be incorporated into tool design we, present Atlantis, a tool designed to assist software security engineers performing program exploitability analysis [2].

Exploitability analysis is the process of determining if a given program may be susceptible to exploitation. One way of determining if a program may have a hidden vulnerability is to: attempt to make the program crash (through a process called fuzzing [9]), trace the program (either by instrumention or an external tracer), and then analyze the resulting execution traces. While this

process can be somewhat automated, assessing the actual exploitability of a crash and performing root cause analysis requires a great deal of human reasoning and manual analysis of the very large trace files generated.

Atlantis is an integrated assembly trace analysis environment designed to assist security engineers perform and manage this analysis. Atlantis was developed in collaboration with software security engineers to meet their requirements, and its design was informed by the work processes summarized in 3.1. Here we reuse these processes to structure our discussion of the features Atlantis offers.

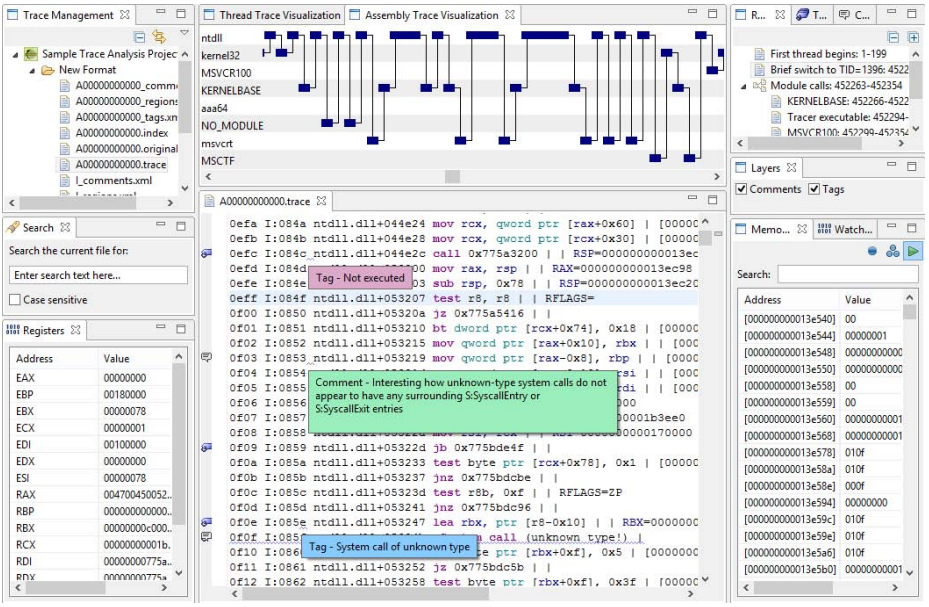


Fig. 1. Atlantis

5.1 Analyzing

With the advent of new tracing technologies (e.g. BitBlaze [7], Pin [5]) we believe analyzing trace files will become the primary task performed by engineers when conducting exploitability analysis of a program. Currently, engineers rely on off-the-shelf text editors and file comparison tools to analyze traces. Atlantis (Figure 1) improves on these tools by providing three customized and linked views: a dedicated Trace Text Viewer, a Trace Visualization View, and a Memory State View which work in concert to allow the engineer to perform their analysis.

1. The **Trace Text Viewer** is a simple text viewer and comparison tool but improves on off-the-shelf solutions with features like very large file support,

fast search, trace-specific syntax highlighting, and memory reference highlighting.
2. The **Trace Visualization View** provides engineers with a high-level representation of the trace. It was designed to help navigate very large traces and to provide a visual overview of the entire trace under study.
3. The **Memory State View** (using an innovative indexing approach) allows an engineer to reconstruct the entire memory state of the program under study at any point in the execution trace, in real time.

5.2 Documenting and Reporting

Documenting trace files (like documenting source code or disassembly) is a similarly crucial part of exploitability analysis. Engineers document traces both to support their analysis activities and to capture and share their findings with other engineers. Atlantis supports this process by providing rich annotation features, allowing engineers to attach comments and tags to locations within the trace.

The Comments View and Tags View provide a way for users to quickly record hypotheses as they traverse the trace. Building on previous work on tagging in software development [8], tags allow a user to annotate a particular line and column (or entire sections) in the trace. There can be multiple occurrences of a tag and using the Tags View, the user can navigate between all occurrences of a tag. Tags can also be grouped into different sets and labeled. Comments function in a similar way but are unique and allow a user to express more complex ideas about a particular location or section of the trace.

Unlike traditional source code editors, where comments and tags are expressed in-line with the source, in Atlantis, comments and tags are displayed in a separate UI layer floating above the Trace Text Viewer and are stored in separate files. This allows the user to selectively display only particular groups of comments and tags. For example, a user analyzing a trace might have different comment groups for different features they are investigating in the trace. Comment and tag layering allows a user to quickly show or hide all comments from one or both of those features.

5.3 Transferring Knowledge

Exploitability analysis offers a lot of opportunity for engineers to collaborate when analyzing traces. Atlantis supports collaboration between engineers by storing annotations separate from traces, and by making it easier for them to share and manage trace annotations. This has multiple benefits.

1. Engineers don't have to share the trace files themselves, but rather just the annotation files. This can be a significant benefit due to the large size of the trace files.
2. As the annotation files are simple xml files, engineers can place them under version control, allowing them to version and share annotations with other engineers in an organized and traceable fashion.

3. If multiple engineers are analyzing a trace collaboratively, they can easily merge annotations from other engineers into their own annotations and then re-export their annotations to be shared with the group.

5.4 Articulating Work

When performing exploitability analysis, engineers will typically not be working with just a single trace, but rather a set of multiple types of traces. For example, along with 'failing' traces that result from program crashes (and which may demonstrate an exploit), engineers often want to analyze and compare 'passing' traces (traces which demonstrated correct operation of the program). The Atlantis Project Management View provides engineers a mechanism for organizing their exploitability analysis of a program into a project structure (including trace and annotation files) and to share that project structure with other engineers through version control. This allows engineers to treat the exploitability analysis of a program as a coherent, standardized entity in itself, rather than as a disparate collection of trace files and documentation.

6 Conclusion and Future Work

The work setting of reverse engineers tasked with security-related issues, such as the dissection of malware or the decryption of encrypted file systems, is unique. Web resources are often unavailable because work has to be performed offline, files can rarely be shared to avoid infecting co-workers with malware or because information is classified, time pressure is immense, and tool support is limited.

In this paper we presented an overview of an exploratory study we conducted [11] to gain an understanding of the work done by security reverse engineers and to understand their processes, tools, artifacts, challenges, and needs. We also reported on Atlantis, a tool that attempts to incorporate the findings of that study and is designed to assist software security engineers with identifying potentially exploitable programs based on analysis of their execution traces. Reverse engineering in a security context is a fast-changing environment. New tools and approaches have to be learned on the spot as hackers and organized cyber groups constantly create new security threats with implications for national security. Future work lies in addressing the challenges that we have identified with improved tools and processes, and in studying their usefulness in the unique work environment of security reverse engineers.

Acknowledgments. We wish to thank the participants in this study, and Cassandra Petrachenko for her feedback on this paper. This research is funded through NSERC grant DNDPJ 380607-09 and DRDC Valcartier.

References

1. Choo, K.K.: Organised crime groups in cyberspace: a typology. Trends in Organized Crime 11, 270–295 (2008)
2. Cleary, B., Painchaud, F., Chan, L., Storey, M.A., Salois, M.: Atlantis - assembly trace analysis environment. In: IEEE 19th Working Conference on Reverse Engineering, WCRE 2012 (2012)
3. Cohen, F.: Computer viruses: Theory and experiments. Computers & Security 6(1), 22–35 (1987)
4. Gerson, E.M., Star, S.L.: Analyzing due process in the workplace. ACM Transactions on Information Systems 4, 257–270 (1986)
5. Luk, C.K., Cohn, R., Muth, R., Patil, H., Klauser, A., Lowney, G., Wallace, S., Reddi, V.J., Hazelwood, K.: Pin: building customized program analysis tools with dynamic instrumentation. In: Proceedings of the 2005 ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2005, pp. 190–200 (2005)
6. Peterson, T.F.: A History of Hacks and Pranks at MIT. The MIT Press (2011)
7. Song, D., Brumley, D., Caballero, J., Jager, I., Kang, M.G., Liang, Z., Newsome, J., Poosankam, P., Saxena, P.: Bitblaze: A new approach to computer security via binary analysis. In: Proceedings of the 4th International Conference on Information Systems Security (2008)
8. Storey, M.A., Ryall, J., Singer, J., Myers, D., Cheng, L.-T., Muller, M.: How software developers use tagging to support reminding and refinding. IEEE Transactions on Software Engineering 43 (2009)
9. Sutton, M., Greene, A., Amin, P.: Fuzzing: Brute Force Vulnerability Discovery. Addison-Wesley (2007)
10. Symantec: Internet security threat report, vol. 17 (April 2012), `http://bit.ly/15nJXO7` (last access: January 3, 2012)
11. Treude, C., Figueira Filho, F., Storey, M.A., Salois, M.: An exploratory study of software reverse engineering in a security context. In: 18th Working Conference on Reverse Engineering (WCRE 2011), pp. 184–188 (2011)

Brain Biomarkers of Neural Efficiency during Cognitive-Motor Performance: Performing under Pressure

Michelle E. Costanzo[1,2] and Bradley D. Hatfield[1,2,*]

[1] Department of Kinesiology, Cognitive Motor Neuroscience Laboratory, University of Maryland, College Park, MD 20742; USA
[2] Neuroscience and Cognitive Science Program, University of Maryland, College Park, MD 20742, USA
bhatfiel@umd.edu

Abstract. The concept of neural efficiency provides a powerful framework to assess the underlying mechanisms of brain dynamics during cognitive-motor performance. Electroencephalography (EEG) studies have revealed that as cognitive-motor performance improves non-essential brain processes are progressively disengaged resulting in brain dynamics leading to a state of neural efficiency. Multiple factors such as practice, genetics, mental stress, physical fitness and social interaction (team dynamics) can influence such cortical refinements positively or negatively and translate into an enhanced or deteriorated quality of performance. This paper provides a report of brain activity, assessed via fMRI, in a group of athletes who perform well under conditions of mental stress. Better understanding of brain states associated with such groups can enhance the ability to detect and classify adaptive mental states and increase the possibility of employing field-friendly brain monitoring tools such as EEG in ecologically valid situations for assessment of cognitive-motor performance in challenging real-world settings.

Keywords: Neural efficiency, expertise, fMRI, emotion regulation.

1 Introduction

Converging neuroimaging data suggest that experts require less neuronal resources compared to novices to accomplish the same task in their domain of expertise, and that this cortical refinement can be characterized as psychomotor efficiency, which is a special case of neural efficiency that refers to the magnitude of communication or input of non-motor brain activity to motor planning processes during movement preparation and execution [1]. Thus, one of the hallmarks of highly skilled individuals is the ability to perform using minimal effort and refined cortical processing specific to the action demands [2]. Many investigators have employed precision aiming tasks

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 123–132, 2013.

(shooting tasks) to explore this notion of efficiency since these kinds of tasks involve minimal movement artifact, and the advantage of ecological validity (e.g.[3]). This research has consistently revealed that the cerebral cortex reduces its activity during task execution, particularly in the left temporal region (associated with verbal analysis), and is characterized by automaticity of motor control [3]. Collectively, these findings imply a refined recruitment of the essential neural networks required for skilled performance. The opportunity to achieve such an adaptive state of cerebral cortical dynamics can be influenced by numerous factors. Personality characteristics, perceptual or attentional styles, trait anxiety, genetic influence on brain processes (e.g., 5-HTT polymorphic influence on emotional states in response to fear-eliciting stimuli), practice, expertise, and social influence as mediated by team dynamics can all affect cortical dynamics to facilitate refinement of networks or introduce nonessential activity that interferes with refinement and efficiency. The relevance of neural efficiency for military operational environments is that state-sensitive biomarkers, such as EEG, heart rate variability, etc., (individually or collectively) could be used to classify if a human operator is in such an adaptive state.

Importantly for this study, such neural efficiency of brain dynamics can become disrupted by mental stress leading to performance decline under pressure [4, 5]. Traditionally, the relationship between stress and performance can be characterized by the organizing principle of the inverted-U, termed the Yerkes- Dodson law [6]. According to this model, performance varies as a function of the stress activation continuum: with an under-aroused-state resulting in sub-optimal performance (in part due to decrements in attention & lack of engagement); a central zone where stress levels are consistent with behavioral adaptability, optimal performance and psychomotor efficiency and extreme excitation, which can become manifested as anxiety, also resulting in performance decline.

As such, the management of high levels of arousal is critical to the performance of tasks under conditions of mental stress. Anxiety-induced disruption of the central zone of optimal arousal may act to perturb the refined process associated with psychomotor efficiency [6]. Such negative appraisal accompanied by elevated arousal, is typically coupled with increased amygdala activity, which, in turn, influences the thalamus, hypothalamus, striatum, and brainstem areas in addition to numerous sensory and association cortical areas [7], creating neuromotor noise. Thus the regulation of emotion (which can be manifested as anxiety), is critical in determining the quality of cognitive-motor performance.

Nonetheless, some individuals are able to maintain a high level of performance during stressful events and, therefore, demonstrate qualities of stress resilience. Stress resiliency encompasses the ability to adaptively cope with adversity and can be examined at behavioral, psychological, and neural levels [8]. For the purpose of the study we define our stress resilient population as individuals who have a history of successful performance (1) senior varsity athletes in the sport of American football 2) letter award winners 3) who typically play a starting role on the team 4) supported by a partial or full athletic scholarship) under conditions of emotional challenge

(high-level competition). Examination of elite performers (intercollegiate athletes) holds promise for understanding the neural basis for such abilities to adaptively cope with stressful events, and more specifically, elite athletes may be uniquely resilient to stress perturbation through the ability to regulate their emotions. Such a population offers a relevant vehicle with which to examine the impact of stress on human performance and can serve as an analogue to military populations who are also challenged with stress while attempting to maintain adaptive performance (i.e., brain) states. The ability to manage or regulate emotion under conditions of mental stress is critical to the quality of performance under such pressure.

There are numerous strategies through which to engage emotion regulatory brain networks, but one strategy, cognitive reappraisal, is a particularly adaptive means of emotion. Cognitive Reappraisal is a "cognitive-linguistic strategy that alters the trajectory of emotional responses by reformulating the meaning of a situation" p 1, [9], and this results in a decrease in the reported negative emotion [10]. In other words, the result of cognitive reappraisal is that it attenuates negative emotional experience resulting in an enhancement in cognitive control of emotion. This implies it is important to consider not only the stressful event, but the individual's perception of the stressor, to understand how skilled performers maintain consistency under various challenges and during mental stress.

In support of this notion, the dynamics between stress (i.e. anxiety) and performance can be further characterized by the transactional model described by Staal (2004) [11]. Specifically, stress is conceived as the aggregate result of the interpretation of the environmental challenge, as well as the objective challenge. In particular, this model integrates human performance and information processing capacity with the notion of appraisal of threat, controllability, and predictability for understanding how stress affects performance. As such, a key element is the individual's appraisal of the situation. This implies that a great deal of individual variation in the response to the stressor may be a consequence of the perception of the event rather than the actual environmental stressor. Therefore, the perception of the stimulus is essential rather than the objective stimulus and, furthermore, the perception may be highly related the individual's experience (i.e. domain specific).

Consequently, elite athletes may have developed a domain-specific reaction to stressful challenge, which through experience and training, allows them to endogenously regulate their affective response to known stressors and efficiently respond to affective challenge. In summary, the present work examined the neuropsychological processes that may well contribute to a state of psychomotor efficiency under stress. Using elite athletes as a model for a stress-resilient population this study attempted to provide insight into the mental approach these individuals employ to maintain mental stability as they engage in sport-specific challenges. A model of stress resiliency is proposed which is characterized by an economy of affective neural processing and an experience-dependent automaticity of neural processes associated with cognitive reappraisal.

2 Materials and Methods

2.1 Participants

Twenty-five male participants between the ages of 18 and 22 were recruited and of these 13 were football athletes (M=21.46 years; SD=0.776) and 12 were non-athletes (M= 21.08 years; SD=2.19).

The football athletes were 1) senior varsity athletes 2) letter award winners 3) typically play a starting role on the team 4) on a partial or full athletic scholarship. The non-athletes were healthy subjects who never played football at a college level, but reported familiarity with the goal and rules of the sport; this is critical to ensure that all subjects understand the meaning of the negative sport-relevant images. Additional selection criteria included that the subjects must have been (a) native English speakers (b) free of current or past diagnosis of neurological or psychiatric disorders, and (c) MRI compatible (e.g., no metal in body, no tattoos on face, no medicine delivery patch). All subjects gave their written informed consent and all experimental procedures were approved by the University of Maryland Institutional Review Board with proper notification IRB of record for Hyman Subject Research Projects performed at the Georgetown University Center for Functional and Molecular Imaging.

2.2 Stimuli

Negative and neutral images were selected from the International Affective Picture System (IAPS). In addition we developed Sport-Specific (SS) images by searching internet databases (e.g., Google Images) to find images representing unpleasant events experienced during football competition: for example: 1) injuries; 2) embarrassment due to loss (i.e., dejected players); 3) critical coaches. SS images were rated with a valence rating mean of 4.131 and arousal mean rating of 4.824. In turn, IAPS images were selected with matching valence means scores of 4.116 and arousal mean scores of 4.896 to create equivalence between the two image sets.

2.3 Task

Each trial was composed of four events: First, instructions (watch or decrease) appeared centrally for 2 seconds. On “decrease” trials, participants were instructed to engage in cognitive reappraisal and on “watch” trials participants will be instructed simply to look at the image and respond naturally. Second, an aversive or neutral image appeared centrally for 8 seconds. While the image remained on the screen, participants performed the evaluation operations specified by the prior instructional cue. Third, a rating scale appeared immediately after presentation of the image for 4 seconds to determine “How negative do you feel” with a rating from 1 to 5 (1 not at all, 3 moderately, 5 extremely). Fourth, the transition task of a fixation cross appeared for 4 seconds in the center of the screen cuing participants to relax until the next trial.

Each subject was cued to passively view or reappraise 48 domain non-specific negative images (24 each) and 48 domain- specific negative images (24 each) in addition to the passive viewing of 24 neutral images during randomly intermixed trials over 4 MRI scanning runs. Each image was shown only once for a given participant.

2.4 Imaging Parameters and Data Analysis

Functional and structural magnetic imaging data were acquired on a 3TSiemens Magnetom Trio system equipped with gradients suitable for echo-planar imaging sequences. Thirty-eight axial slices (3.2 mm thick in plane) were acquired using an echo planar imaging (EPI) pulse interleaved sequence (TR 2000 ms; FOV 205; TE 30ms). The DICOM images imported Statistical Parametric Mapping, SPM5. Slice timing and head motion correction, was followed normalization into MNI format (template EPI.mni). Default SPM5 settings were used to warp volumetric MRIs to fit the standardized template (16 nonlinear iterations), and normalization parameters were applied to subject's functional images. Normalized images were resampled into 2 × 2 × 2 mm voxels and smoothed. Preprocessed images were entered into a General Linear Model in SPM5 that modeled the canonical hemodynamic response function convolved with an 8-second boxcar representing the picture-viewing period. Motion parameters, the instructional cue period, and the rating period were entered into the model as additional regressors. Contrasts were created for each condition relative to the neutral baseline. These individual contrasts were then entered into a Full Factorial design of a 2 x 4 ANOVA Group by Conditions to perform a random-effects group analysis. The Group factor consisted of Athlete and Control and the Condition factor consisting of Cognitive Reappraisal SS, Passive Negative SS, Cognitive Reappraisal IAPS, and Passive Negative IAPS. Whole brain analysis was examined for each group relative to the neutral condition. Region of interest analysis was executed for the Cognitive Reappraisal SS vs Passive Negative SS, Cognitive Reappraisal IAPS vs Passive Negative IAPS in the prefrontal cortex (BA 8, 9, 10, 11, 45, 46, 47, taken from the Wake Forest Pick Atlas indication of Brodmann Areas). All results were FDR corrected for multiple comparisons ($p<0.05$) unless otherwise noted.

3 Results

3.1 Whole Brain Analysis

Whole brain analysis revealed that during the natural response of the athlete group to generalized negative images (IAPS) (relative to the neutral baseline) significant activation occurred in the left dorsolateral prefrontal cortex (DLPFC), left inferior frontal gyrus (IFG), left dorsomedial prefrontal cortex (DMPFC), left ventrolateral prefrontal cortex (VLPFC), bilateral orbitofrontal cortex (OFC), bilateral superior parietal lobule (SPL), right lingual gyrus, bilateral parahippocampal gyrus, bilateral premotor cortex (PMC), right cerebellum, superior temporal gyrus (STG) and right middle temporal gyrus (MTG). In the control group significant activation was observed in the bilateral DLPFC, left DMPFC, left IFG, left VLPFC, the bilateral OFC, the right STG, left

inferior temporal gyrus (ITG) and right middle occipital gyrus (MOG), right anterior cingulated cortex (ACC), bilateral PMC, left SPL, right lentiform nucleaus and right postcentral gyrus.

Whole brain analysis results indicate that during passive viewing of sports-specific (SS) images, the athlete group exhibited significant activation in the left DMPFC, left insula, right lingual gyrus, left DLPFC, bilateral VMPFC, left IFG, right OFC, bilateral PMC, bilateral SPL/precuneous, left postcentral gyrus, bilateral ITG, the right STG, bilateral parahippocampal gyrus, left putamen and left thalamus. During the passive viewing of SS images, the control group exhibited significant activation in the bilateral DLPFC, bilateral VMPFC, right IFG, left OFC, left insula, bilateral STG, left ITG, right lingual gyrus, bilateral parahippocampal gyrus, bilateral PMC, bilateral SPL, left precentral gyrus, and left and right lentiform nucleus.

Cued cognitive reappraisal of generalized negative images (IAPS) resulted in significant activation of the left DLPFC, bilateral VMPFC right VLPFC, bilateral OFC, right lingual gyrus, bilateral premotor cortex, bilateral parahippocampal gyrus, left post central gyrus, right SPL, bilateral ITG, left MTG, bilateral cerebellum, left uncus and left lentiform nucleus in the athlete group. Significant activation in the left DLPFC, left DMPFC, bilateral OFC, bilateral IFG, bilateral PMC, right SPL, left supramarginal gyrus, left amygdala, right MOG, left posterior cingulated, right STG, right ITG, left MTG, and the right cerebellum was observed in the control group during cued reappraisal of IAPS images.

The cued cognitive reappraisal of SS images revealed significant activations in the left DLPFC, left VLPFC right OFC, bilateral PMC, right lingual gyrus, bilateral parahippocampal gyrus, left supramarginal gyrus right postcentral gyrus, bilateral SPL,

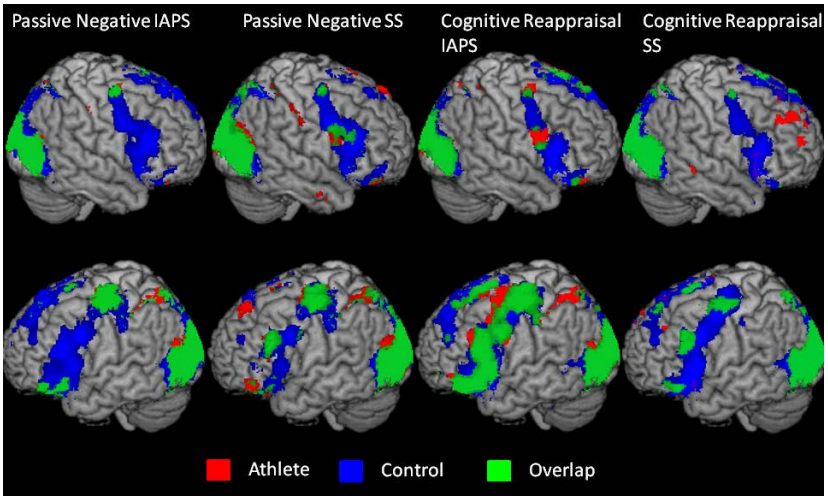


Fig. 1. Results of whole brain analysis. Passive Negative IAPS, Passive Negative SS, Cognitive Reappraisal IAPS, and Cognitive Reappraisal SS contrasts are relative to the neutral baseline. The red indicates the unique activation for the athlete group, the blue indicates the unique activation of the control group, and the green indicates regions where both groups showed activation (overlap), $p<0.05$, FDR corrected.

left STG, left MTG, left lentiform nucleus and left cerebellum in the athlete group. Activation was observed in the control group in the bilateral DLPFC, left IFG bilateral VLPFC right VMPFC bilateral medial OFC right cuneus, left parahippocampal gyrus, bilateral PMC, left MTG bilateral SPL, bilateral lentiform nucleus, bilateral STG, right motor cortex left posterior cingulate and bilateral insula.

3.2 Region of Interest Analysis

The region of interest analysis of the Cognitive Reappraisal of IAPS images and the Passive Response to IAPS images indicated activation in the left DLPFC, the bilateral DMPFC, bilateral VLPFC, right VMPFC and left IFG ($p<0.05$, uncorrected) in the athlete group (Figure 2). Direct comparisons within the IAPS image set between Cognitive Reappraisal and Passive Negative revealed that during cued cognitive reappraisal the left IFG ($p<0.05$, uncorrected) was active in the control group (Figure 2).

The region of interest analysis revealed that no difference ($p<0.05$, uncorrected) was detected during the Cognitive Reappraisal SS- Passive Negative SS contrast in the prefrontal for the athletes (Figure 2). Direct comparisons of Cognitive Reappraisal SS - Passive Negative SS revealed greater activation in the left DMPFC (BA 8), left IFG (BA 47), and right IFG ($p<0.05$, uncorrected) in the control group (Figure 2).

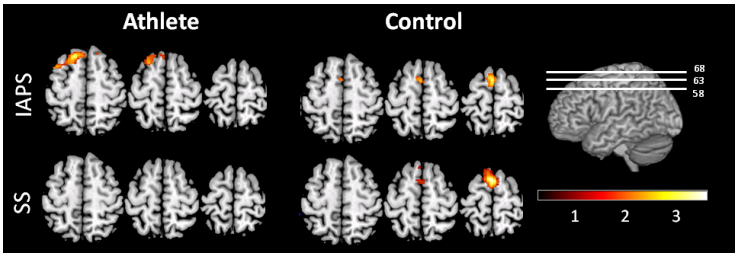


Fig. 2. Axial slices from prefrontal cortex region of interest analysis for the Athlete (left panel) and Control (right panel) groups. Contrast: Cognitive Reappraisal - Passive Viewing. IAPS. Generalized Negative Images. SS. Sports-Specific Negative Images. Slice numbers and t-score color bar is provided ($p<0.05$, uncorrected).

4 Discussion and Conclusion

There is a robust relationship between one's emotional state and the ability to effectively perform cognitive-motor skills. Elite performers must balance competing task demands such as physical requirements (dexterity, force), physiological recovery (metabolic rate, body temperature), psychological focus (memory, decision making), etc. during high levels of performance [12]. Critical to the orchestration of adaptive responses to the challenge of competitive sport, is the management of the emotional component of the task. Mental stress can lead to detrimental outcomes like state anxiety, burnout, exhaustion, strain, and tension, but it can also evoke adaptations such as hardiness, resilience and resistance [13]. Thus, these divergent outcomes must be

explained not only in terms of the nature of the stressor but also in terms of the individual's perception of the challenge. Our data revealed a generalized neural processing efficiency during affective challenge (Figure 1) in which elite athletes are less perturbed by mental stress and suggests this may be a critical quality contributing to their stress resilience. When examining the specific patterning of neural processing during the natural response of the elite athletes to stressful challenges, our data show that they demonstrate similar neural processes to those used during cognitive reappraisal, but this is only within their domain of expertise (Figure 2). Thus, the confluence of experience based-factors such as controllability, emotional coping strategies, motivational efforts, trait/state anxiety and individual personality, in addition to the qualities of the objective stressor, cumulatively interact to produce the stress response.

The disruptive effects of stress on human performance can be classified as a loss of neural processing efficiency [1] leading to hyperactivity of non-essential brain regions that interfere with the cognitive-motor task demands. Conceptualized as "neuromotor noise," [14] this process affects cortical arousal and redistributes processing resources away from those dedicated to the goal-directed behavior. The loss of neural processing efficiency caused by stress-induced neuromotor noise may explain the phenomenon of "choking" or performance decline under pressure [4, 5]. However, elite-level athletes are typically resilient to such stress perturbation, enabling them to maintain a high level of performance during stressful conditions. The whole brain analysis result, which revealed more focused brain activity in the athletes during all conditions, suggests that neural efficiency in the motor domain as reported in the literature [1] extends to the emotional domain (Figure 1). This, in turn, would promote an overall refinement of cortical activity necessary for successful performance under mental stress and allow for a greater capacity to handle stressful events (i.e., less neuromotor noise).

Interestingly, the elite athletes demonstrate efficiency during both specific (sport-specific) and generalized (IAPS) challenge. On speculation, this pattern may be a consequence of repeated exposure to competitive stress, which can lead to active coping strategies that would translate to an ubiquitous planning and problem solving approach to challenge [8]. Our results also support efficiency in brain regions sensitive to social competence and understanding, which may promote adaptive neural processing mediated by oxytocin (reduces fear response) [8]. In addition physical fitness is associated with altered behavioral and neuromodulator responses to stressors (e.g.[15]). Lastly, genetic factors could also contribute to adaptive responses to stress by way of mediating reward circuits and protecting against depression [16] and trait disposition to anxiety [17]. Our present design cannot address the speculations identified here, but we examined one specific element of stress resiliency, cognitive reappraisal.

Cognitive reappraisal is a cognitive-linguistic strategy that changes the trajectory of emotional responses by reformulating the meaning of a situation such that negative affect experience is reduced [9]. Thus cognitive reappraisal serves 1) as a means for understanding the qualities that contribute to the unique features of stress resilient population compared to a representative sample population and 2) a critical reference

for understanding what stress resilient individuals do when responding naturally to stressful events. Neuroimaging studies have examined this cognitive approach to mental stress and have revealed that frontally mediated executive processes act to manage the response of the amygdala (central to emotional processing)[9].

We examined if those who have demonstrated stress resilience (superior performance under pressure) exhibit such a specific pattern of neural responses characterized by this adaptive emotion regulatory strategy (cognitive reappraisal) in the prefrontal cortex. In addition, as stated earlier, the transactional model [11] predicts a high degree of specificity of the stress response based on an individual's perception and appraisal of the stressful event. Consequently, an athlete may have developed through experience and training a domain-specific reaction to stressful challenge, which allows them to endogenously regulate their affective response to familiar stressors. The region of interest analysis between the sport specific conditions (cued cognitive reappraisal and passive viewing of negative sport-specific images) indicates that through experience, these individuals automatically engage in mental transformation of an emotional event such that the negative consequences are attenuated, (i.e. they appear to endogenously engage in cognitive reappraisal) (Figure 2). This equivalence of processing (no difference during SS in athletes) between the natural response to mental stress and cued cognitive reappraisal is lost during the generalized negative events (IAPS images). Although this work was based on MR imaging future work can employ EEG/fNIRS, which are more appropriate in operational environments. A longer term goal is to develop applications for brain monitoring of emotion level and regulation in the field. Such an approach could be applied to military personnel for monitoring during combat situations for the purpose of stress management, altering workload based on one's state as well as for solider selection for special units if robust profiles emerge.

The results suggest that skilled performers who excel during competitive stress engage in cognitive regulation in their domain of expertise, decreasing physiological arousal thereby enabling them to sustain elevated performance. This specificity suggests that emotion regulation promotes refinement of brain activity resulting in an optimal state for effective task execution particularly under conditions of known stressful challenge (i.e., sport competition). By investigating a stress resilient population (elite athletes), this study provides an assessment of the postulated dynamic between cognitive (prefrontal) and affective (limbic) brain networks as related to skilled motor performance. What emerges is a generalized neural efficiency that appears to be a quality of resiliency to promote a mental state where neuromotor noise is attenuated. However a specific element of resiliency (i.e., automaticity of cognitive reappraisal) is dependent on experience. In the context of performance, cognitive reappraisal, through prefrontal regulation of the arousal, may maintain an adaptive level of arousal to promote a state of psychomotor efficiency during mental stress. The establishment of this protocol as an effective means through which to probe the emotion regulatory processes in elite groups, holds promise to facilitate more tactical psychological interventions that aid in motor performance.

References

1. Hatfield, B.D., Kerick, S.E.: The Psychology of Superior Sport Performance: A Cognitive and Affective Neuroscience Perspective. In: Tenenbaum, G., Eklund, R.C. (eds.) Handbook of Sport Psychology, pp. 84–109. John Wiley & Sons, Inc. (2007)
2. Del Percio, C., Rossini, P.M., Marzano, N., Iacoboni, M., Infarinato, F., Aschieri, P., Lino, A., Fiore, A., Toran, G., Babiloni, C., Eusebi, F.: Is there a "neural efficiency" in athletes? A high-resolution EEG study. Neuroimage 42, 1544–1553 (2008)
3. Haufler, A.J., Spalding, T.W., Santa Maria, D.L., Hatfield, B.D.: Neuro-cognitive activity during a self-paced visuospatial task: comparative EEG profiles in marksmen and novice shooters. Biol. Psychol. 53, 131–160 (2000)
4. Beilock, S.L.: Choke: What the Secrets of the Brain Reveal about Getting it Right When You Have. Free Press, New York (2010)
5. Beilock, S.L., Carr, T.H.: On the fragility of skilled performance: what governs choking under pressure? J. Exp. Psychol. Gen. 130, 701–725 (2001)
6. Hancock, P.A., Szalma, J.L.: Performance Under Stress. Ashgate Publishing, Burlington (2008)
7. Haines, D.E.: Fundamental Neuroscience for Basic and Clinical Applications. Churchill Livingstone Elsevier, Philadelphia (2006)
8. Feder, A., Nestler, E.J., Charney, D.S.: Psychobiology and molecular genetics of resilience. Nat. Rev. Neurosci. 10, 446–457 (2009)
9. Goldin, P.R., McRae, K., Ramel, W., Gross, J.J.: The neural bases of emotion regulation: reappraisal and suppression of negative emotion. Biol. Psychiatry. 63, 577–586 (2008)
10. Wager, T.D., Davidson, M.L., Hughes, B.L., Lindquist, M.A., Ochsner, K.N.: Prefrontal-subcortical pathways mediating successful emotion regulation. Neuron 59, 1037–1050 (2008)
11. Staal, M.A.: Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework. National Technical Information Service, N.C.f.A. Information (2004)
12. Andre, A.: The value of workload in the design and analysis of consumer products. In: Hancock, P.A., Desmond, P.A. (eds.) Stress, Workload, and Fatigue, pp. 373–382. L. Erlbaum, Mahwah (2001)
13. Tepas, D.I., Price, J.M.: What is stress and what is fatigue? In: Hancock, P.A., Desmond, P.A. (eds.) Stress, Workload, and Fatigue. L. Erlbaum, Mahwah (2001)
14. Van Galen, G.P., Van Huygevoort, M.: Error, stress and the role of neuromotor noise in space oriented behaviour. Biol. Psychol. 51, 151–171 (2000)
15. Dishman, R.K.: Brain monoamines, exercise, and behavioral stress: animal models. Med. Sci. Sports Exerc. 29, 63–74 (1997)
16. Vialou, V., Robison, A.J., Laplant, Q.C., Covington, H.E., Dietz III, D.M., Ohnishi, Y.N., Mouzon, E., Rush, A.J., Watts III, E.L., Wallace, D.L., Iñiguez, S.D., Ohnishi, Y.H., Steiner, M.A., Warren, B.L., Krishnan, V., Bolaños, C.A., Neve, R.L., Ghose, S., Berton, O., Tamminga, C.A., Nestler, E.J.: DeltaFosB in brain reward circuits mediates resilience to stress and antidepressant responses. Nat. Neurosci. 13, 745–752 (2010)
17. Canli, T., Ferri, J., Duman, E.A.: Genetics of emotion regulation. Neuroscience 164, 43–54 (2009)

The Geometry of Behavioral and Brain Dynamics in Team Coordination

Silke Dodel[1], Emmanuelle Tognoli[1], and J.A. Scott Kelso[1,2]

[1] Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, FL 33431, USA
[2] Intelligent Systems Research Center, University of Ulster, Derry, N. Ireland
dodel@ccs.fau.edu

Abstract. Performing a task as a team requires that team members mutually coordinate their actions. It is this coordination that distinguishes the performance of a team from the same actions performed independently. Here we set out to identify signatures of team coordination in behavioral and brain dynamics. We use dual electroencephalography (EEG) to measure brain dynamics of dyadic teams performing a virtual room clearing task. Such complex tasks often exhibit high variability of behavioral and brain dynamics. Although such variability is often considered to impede identification of the behavior or brain dynamics of interest here we present a conceptual and empirical framework which explains variability in geometrical terms and classifies its sources into those that are detrimental and non-detrimental to performing the task at hand. Using our framework we found that behaviorally team coordination is reflected in terms of role dependent behavior. Furthermore we identified a low-dimensional subspace of the brain dynamics in the frequency domain which is specific for team behavior and correlated with successful team coordination. Moreover, successful team coordination was positively correlated with the inter- but not intra-brain coherence in the gamma band. Our results hence indicate that successful team coordination is associated with increased team cognition, particularly readiness to engage in the task.

1 Introduction

Many tasks in real life are best accomplished when a group of people acts as a team. Examples include lifting heavy weights, hunting, police or military operations, and team sports, but also some instances of abstract problem solving. Performing a task as a team requires that team members mutually coordinate their actions. It is this coordination that distinguishes the performance of a team from the same actions performed independently by multiple subjects. While multiple studies investigate team cognition from behavioral measures [1,2,3,4,5,6,7,8,9,10], it is currently unknown how team cognition is reflected in the brain dynamics of the team members. Studies of brain activity during social interaction - but not team cognition - have found brain rhythms associated with social coordination [11], inter-brain coherence [12,13], inter-brain

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 133–142, 2013.

Granger causality [14], and joint brain networks [15]. The findings of the few extant studies about brain activity during team tasks include indications of inter-brain functional connectivity in partners of the same team in a card game [16], dimensionality of brain dynamics being affected by team expertise and task difficulty [17], and classification of EEG engagement patterns of individual team members [18].

Here we use dual electroencephalography (EEG) of dyadic teams who perform a virtual room clearing task to identify neuromarkers of team coordination. The task we use is a virtualization of one of the most extreme forms of team coordination, namely when members' survival and safety depend upon efficient team interactions. During such tasks cognitive and social processes have to be coordinated in a context-dependent fashion. The task is ecologically valid and highly dynamic with well defined behavior in which subjects dynamically engage in and disengage from team coordination. Team coordination can be studied in a meaningful way only in ecologically valid tasks. Ecologically valid tasks however are often highly dynamic and individuals performing such tasks exhibit a high variability of behavioral and brain dynamics. Here we present a conceptual framework which explains this variability in geometrical terms and classifies variability into detrimental and non-detrimental to the task at hand. Our framework provides a unifying theoretical account for tasks with multiple degrees of freedom and hence is particularly suited to guide the analysis of team tasks.

The paper is organized as follows: In section 2.1 we describe the experiment, in section 2.2 the conceptual framework for analyzing behavioral and brain dynamics in complex tasks is introduced, and in section 3 we present the results of this framework, followed by a discussion in section 4.

2 Materials and Methods

2.1 Virtual Room Clearing Task

For our study we devised a virtual room clearing task by creating a video game in which team members of a dyadic team work together to detect and eliminate enemies as they jointly progress through buildings in a hostile urban environment, with the shared goal of clearing a virtual room from threats. The video game was designed to retain the essence of key behavioral, perceptual, cognitive, social and attentional processes that occur in successful team work. Subjects shared the same top down perspective of their virtual environment while controlling their avatar's position and direction of gaze, as they navigated through a series of 32 buildings each composed of 5 successive rooms, with their virtual environment becoming visible upon the avatars' spatial exploration. Dual-EEG was recorded by using two 60 channel EEG caps and a sampling rate of 1 kHz. The experimental setup is described in more detail in [19]. Figure 1(a) shows an example trajectory of an avatar dyad through a building. In each room clearing task one subject assumes the role of the leader and the other the role of the follower. The subjects assume the roles of leader and follower in a self-organized manner and renegotiate their roles implicitly by the behavior of their avatars

after completion of each room clearing trial. There are two valid entry patterns, the leader button-hook entry and the leader cross-over entry (cf. Figure 1(b)). Due to this degeneracy, the task is endowed with intrinsic variability. On the day before the experiment, subjects received training for one hour to master basic concepts of room clearing. Although in our study the room clearing task is virtual, the behavior of the avatars closely resembles the behavior of subjects performing an actual room clearing task (compare Fig. 1(c) to Fig. 3, left).

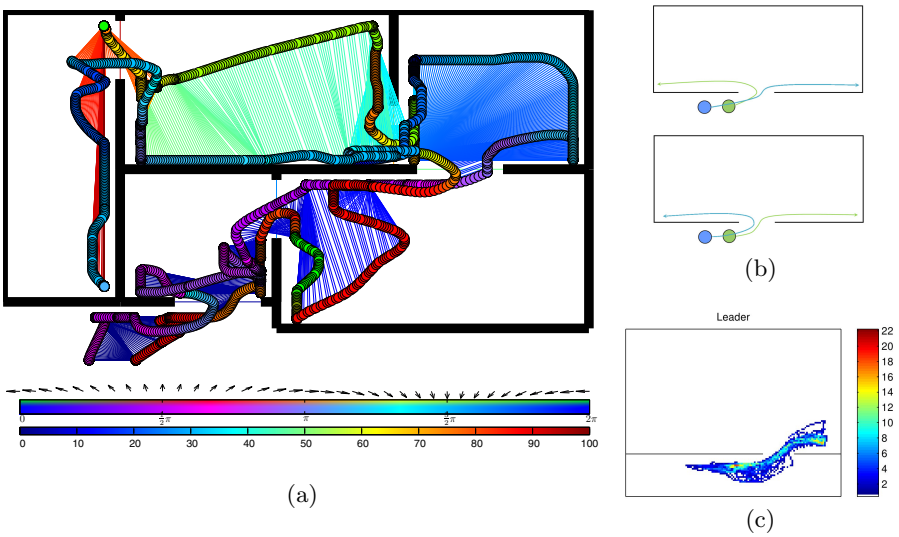


Fig. 1. (a) Trajectories of the avatar dyad through the five rooms of one of the buildings. Color of the beams between team member positions: percentage of trajectory completed (from blue 0% to red 100%). Color of the circles: Gaze direction. Green subject: green 0°, red 120°, yellow 240°. Blue subject: blue 0°, magenta 120°, cyan 240°. (b) Two possible team entry patterns: leader button-hook (top) and leader cross-over (bottom). The follower executes the complementary entry pattern, respectively. (c) Spatial probability density of leader trajectories in a behavioral (not virtual) room clearing task. Values indicate the percentage of trajectories crossing a given region. Based on data from [8].

Each room clearing trial proceeds as follows: The follower aligns behind the leader. The follower gives a ready signal, conveyed through touch 'tap' in the real situation, and through a vibration of the Xbox controller in the experiment (verbal communication was not allowed to avoid EEG artifacts) upon which the leader initiates motion of his avatar and is followed by the avatar of the follower. Then both subjects move their avatars to the door as close as possible and the leader presses a button to open the door. The leader enters the room, followed as closely as possible by the follower. Then both leader and follower independently move to their respective corners of dominance, which are the corners to the left

and right from the door. Both corners of dominance must be covered, hence the follower needs to coordinate his entry pattern with that of the leader such as to move to the opposite corner as the leader. After arriving in their corners of dominance, subjects had instruction to engage in gaze interlock, i.e. the overlap of their view cones. The gaze interlock ended the trial and the subjects moved to the next stacking point. From the description above, each room clearing trial can be divided into three main events: First, the coordination build-up while stacking in front of the room, second, coordinated behavior when moving to the door and entering the room, and third, breaking and rebuilding coordinated behavior.

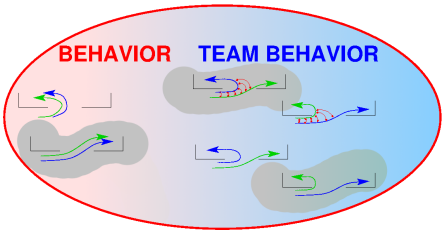
2.2 Geometrical Description of Behavioral and Brain Dynamics during Complex Tasks

Detrimental and Non-detrimental Variability. Ecologically valid tasks are often complex and have a high number of degrees of freedom which leads to behavioral variability because there are multiple valid ways in which these tasks can be executed (cf. Fig. 1(b)). Such behavioral variability is non-detrimental to task execution and actually may improve its performance metrics (e.g. speed, survival, ...). However, other variability may be detrimental, representing a deviation from ideal task execution. Furthermore in team tasks the behavior of each team member is influenced by the behavior of the other team members. This gives rise to particular behavioral patterns which are not present in tasks performed by a single subject. In addition to behavioral variability there is also variability in the brain dynamics associated with task execution. Neural processes also exhibit degeneracy, and hence a single behavior may be supported by multiple neural processes.

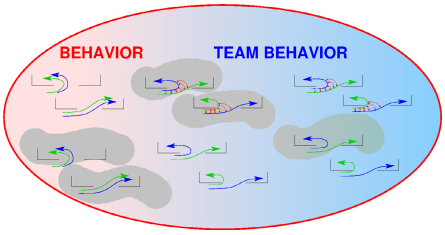
To guide the choice and development of analysis methods we devised a geometrical framework for the description of complex tasks. This allows to translate properties of the behavioral and brain dynamics into geometrical properties and makes available geometrical tools for the analysis of dynamics in brain and behavior. An implicit geometrical perspective is already adopted in many standard multivariate analysis methods such as principal component analysis or clustering algorithms. Our framework is novel however, because it offers a conceptualization of all behavioral aspects of a task and its associated brain dynamics in terms of geometry and furthermore provides a geometrical account for inter subject coordination.

Behavioral and Brain Manifolds. Behavioral and brain dynamics during a complex task can be described geometrically as a trajectory in a phase space. Each dimension of the phase space represents a variable of the dynamics. The choice of the phase space is not unique. Different phase spaces hence provide different "windows" to the problem. The brain dynamics in each trial represents a trajectory in phase space. The ensemble of all possible trajectories constitutes a geometrical object which we refer to as manifold. The term "manifold" is chosen

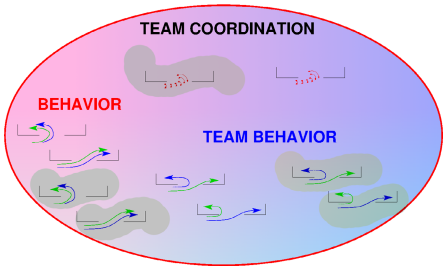
to indicate that the geometrical object can have a non-trivial shape and alludes to the Uncontrolled Manifold in movement sciences [20]. Here we extend the uncontrolled manifold concept to team behavior and brain dynamics. Figure 2 shows the manifold concept for behavioral and brain dynamics, respectively. A trajectory on the behavioral manifold contains information about the team behavior in a given trial, and a trajectory in the brain manifold represents information about the associated brain dynamics. Note that unless stated otherwise, the term "trajectory" always refers to a trajectory on the manifold rather than to a physical trajectory through the virtual rooms. The shape of the manifolds is governed by three main factors: task constraints, team coordination and task performance. Task constraints affect the global shape of the manifold while team coordination and task performance delineate submanifolds.



(a) Behavioral manifold



(b) Brain manifold (team coordination: *modulation* of neural processes)



(c) Brain manifold (team coordination: *separate* neural processes)

Fig. 2. Illustration of the manifold concept in phase spaces of behavioral and brain dynamics, respectively. (a) Behavioral manifold. (b) Brain manifold in which team coordination is supported by modulation of neural processes underlying behavior. (c) Brain manifold in which coordination is supported by *separate* neural processes. The cartoons on the brain manifold stand for the behavior that the neural processes support. Red connections illustrate team coordination. Due to degeneracy, the same behavior may be supported by different neural processes. In all manifolds, team performance increases from left to right, and team coordination from bottom to top. The shaded areas in the manifolds illustrate a probability measure on the manifold.

Team Subspace. A manifold can be approximated at any given point by its tangent space at this point. Here we use singular value decomposition (SVD) of

local time-frequency windows to approximate the tangent spaces of the manifolds. Neuromarkers are properties of brain dynamics that are related to a certain behavior, which in our case is team coordination. For neuromarkers to be identifiable, certain properties of brain dynamics during the behavior in question need to be sufficiently consistent. Consistency, however, implies low variability and hence we are confronted with the challenge to identify consistent patterns from highly variable brain dynamics. To identify consistent brain dynamics related to team behavior, yet account for the variability associated with neural and behavioral degeneracy, we compute the intersection of the tangent spaces over different trials and project the brain signals into this subspace which we refer to as team subspace. The subspace represents the submanifold of the brain manifold which is related to team behavior and the projected brain signals correspond to trajectories on this submanifold. We determined the intersection of the tangent spaces by computing the principal angles between them using a method proposed in [17]. The dimensionality of the team subspace was determined by the number of dimensions for which $\cos\alpha > 0.8$ where α is a principal angle.

Measure of Team Coordination. When the subjects perform the virtual room clearing task, task initiation is a key event of team coordination. The task is initiated when the follower taps the leader to signal his readiness to start upon which the leader starts to move. A short duration of this time interval indicates that both subjects were simultaneously ready to start the task. Here we use the duration of task initiation as global measure of team coordination of a trial. Short durations thereby indicate successful team coordination.

3 Results

Behavioral Signatures of Team Coordination. Our study focusses on brain signatures of team coordination, however, we found that behavioral signatures of team coordination were present even on the level of single subject behavior. More specifically, we found that the entry patterns of leaders and followers exhibited role-dependent differences. Figure 3 shows the behavioral variability over trials of cross-over entries performed by a leader and a follower, respectively. The entry pattern of the leader has much less variability than that of the follower. There are two regions of high concentration of leader entry trajectories, one before and one after the entry. Interestingly, this pattern is also found in the leader trajectory during actual room clearing (cf. Fig. 1(c)). The entry pattern of the follower is much more diffuse with only one area of slightly higher concentration of entry trajectories. This indicates that the follower is coordinating with the leader. The coordination of the follower with the leader has the effect of enhancing the variability of his own behavior, since his behavior is affected by both external task constraints and the behavior of the leader, while the leader's behavior is primarily affected by external task constraints.

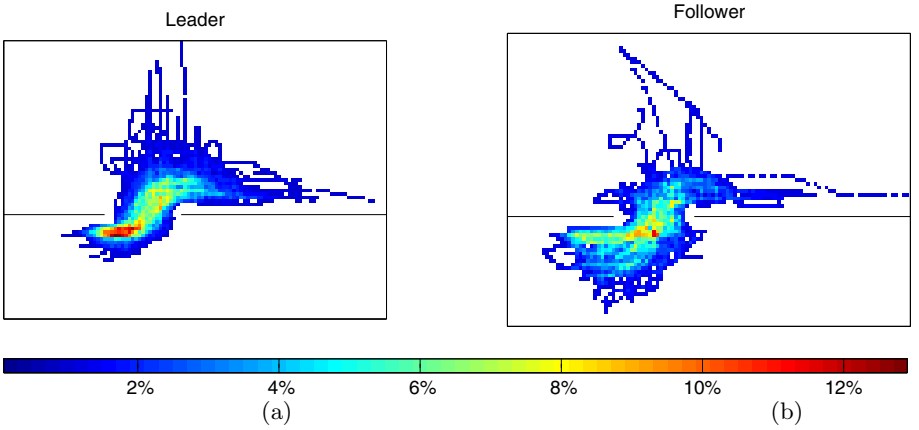


Fig. 3. Behavioral variability of a cross-over entry of a leader and a follower. Spatial probability density of the trajectories. Values indicate the percentage of trajectories crossing a given region. (a) Leader. (b) Follower.

Brain Signatures of Team Coordination. We applied our geometrical framework (cf. Materials and Methods) to the dual EEG data from the virtual room clearing task and determined a brain manifold in the frequency domain, using as phase space variables the real and imaginary parts of a wavelet transform (complex Morlet, 3 peaks) of the signal of all electrodes. We restricted our analysis to the frequency band of 15-40 Hz (β and low and medium γ band), because the behaviors of interest like signaling readiness by the follower, movement initiation by the leader, and entry coordination between the two subjects occur at time scales of 200 ms or less, corresponding to 3-25 cycles in the chosen frequency band. Here we report results from the task initiation interval, which is the interval between the tap performed by the follower to signal readiness and the movement onset of the leader. This interval has a duration of 250 ms - 1000 ms with an average around 400 ms. Figure 4(a) shows the spatio-temporal patterns of the wavelet powers in the frequencies 15-40 Hz during the first 250 ms of the task initiation interval, averaged over trials for one dyadic team. We found consistent low-dimensional team subspaces (cf. Materials and Methods) for each dyadic team, with dimensionalities of 8-11. Neuromarkers of team behavior were identified by the portion of the brain signals that lay within the team subspace (cf. Fig. 4(b)). We found the mean relative wavelet power in the team subspaces to be significantly positively correlated with successful team coordination as measured by the duration of the task initiation interval (cf. Materials and Methods). This supports the notion of a low-dimensional subspace of joint brain activity which is related to team coordination. Moreover we found that inter-brain coherence, but not intra-brain coherence in the low and medium γ bands (30-40 Hz) during task initiation was significantly positively correlated with

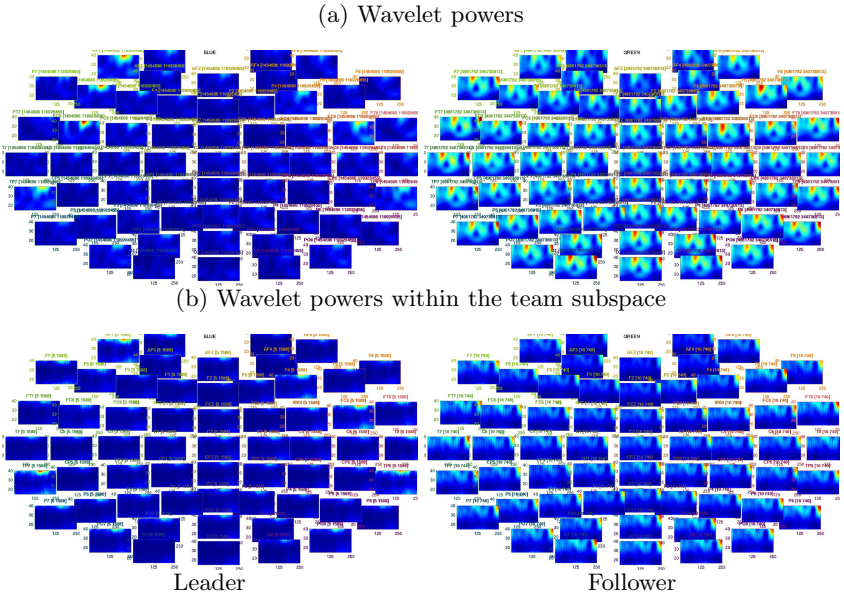


Fig. 4. Mean wavelet powers of leader (left column) and follower (right column) (a) before and (b) after projection onto the team subspace

successful team coordination as well. With inter-brain coherence and spatio-temporal wavelet powers in a team subspace, we have identified two neuromarkers of team coordination.

4 Discussion

Social interaction is a crucial part of human life. Yet the neuroscience of social interaction is only in its infancy. To identify neuromarkers of team coordination we have performed a dual EEG study of dyadic teams performing a virtual room clearing task. Ecologically valid tasks like this represent a particular challenge to the analysis because of their high inherent variability. We have developed a geometrical framework that provides a unifying theoretical account for ecologically valid tasks and is particularly suited to guide the analysis of team tasks. In this framework behavioral and brain dynamics are interpreted as evolving along trajectories on a manifold in a high-dimensional phase space. The geometry of the manifold is determined by multiple factors, such as task constraints, team performance and team coordination. The manifold depends on the choice of the phase space. Here we analyzed the data in the wavelet domain, which lends itself to two natural descriptors: wavelet power and coherence, representing amplitude and phase information, respectively.

While our focus was on neuromarkers, we found behavioral signatures of team coordination even on the single subject level in terms of role dependent behavior.

Their pattern indicates that the follower is coordinating his behavior with that of the leader akin to an object attached to the leader. On the level of the brain, successful team coordination was associated with an increased proportion of the subjects' brain signals evolving in a low-dimensional team subspace. Moreover, we found successful team coordination to be positively correlated with inter- but not intra-brain coherence in the γ band during task initiation. This implies that inter-brain coherence is higher when both subjects are ready to engage in the task. Since task initiation is associated with information transfer, inter-brain coherence might be a signature of a state of mind of both subjects which enables or facilitates information transfer. While it has been argued that coherently oscillating neuronal groups facilitate efficient interaction within a brain [21], there is currently no extablished model of how inter-brain coherence could be mediated [22]. A possible interpretation of our finding is that inter-brain coherence is related to the subjects' receptiveness of their mutual behavior. However, a more detailed analysis of the spatio-temporal patterns of intra- and inter-brain coherence in relation to the subjects' behavior is needed to test this hypothesis.

Behavioral variability is brought about by the degrees of freedom of the task as well as differences in task performance. Variability in brain dynamics is due to behavioral variability and due to the degeneracy of brain processes. It is well known - although rarely explicitly acknowledged - that multiple different brain processes may support identical behavior [23]. Our framework represents a departure from the quest for single neural mechanisms underlying given behaviors and explicitly acknowledges the degeneracy and highly synergistic nature of brain processes.

References

1. Fiore, S.M., Salas, E., Cuevas, H.M., Bowers, C.A.: Distributed coordination space: Toward a theory of distributed team process and performance. Theoretical Issues in Ergonomics Science 4(3-4), 340–364 (2003)
2. Cooke, N.J., Gorman, J.C., Duran, J.L., Taylor, A.R.: Team cognition in experienced command-and-control teams. J. Exp. Psychol. Appl. 13(3), 146–157 (2007)
3. Woolley, A.W., Hackman, R.J., Jerde, T.E., Chabris, C.F., Bennett, S.L., Ko sslyn, S.M.: Using brain-based measures to compose teams: How individual capabilities and team collaboration strategies jointly shape perf ormance. Social Neuroscience 2(2), 96–105 (2007)
4. Salas, E., Cooke, N.J., Rosen, M.A.: On teams, teamwork, and team performance: discoveries and developments. Hum Factors 50(3), 540–547 (2008)
5. Bourbousson, J., Sve, C., McGarry, T.: Space-time coordination dynamics in basketball: Part 1. intra- and inter-couplings among player dyads. J. Sports Sci. 28(3), 339–347 (2010)
6. Bourbousson, J., Sve, C., McGarry, T.: Space-time coordination dynamics in basketball: Part 2. the interaction between the two teams. J. Sports Sci. 28(3), 349–358 (2010)
7. DeChurch, L.A., Mesmer-Magnus, J.R.: The cognitive underpinnings of effective teamwork: a meta-analysis. J. Appl. Psychol. 95(1), 32–53 (2010)

8. Dodel, S., Pillai, A., Fink, P., Muth, E., Stripling, R., Schmorrow, D., Cohn, J., Jirsa, V.: Observer-independent dynamical measures of team coordination and performance. Motor Control: Theories, Experiments, and Applications, 72–101 (2010)
9. Gorman, J.C., Amazeen, P.G., Cooke, N.J.: Team coordination dynamics. Nonlinear Dynamics Psychol Life Sci. 14(3), 265–289 (2010)
10. Gorman, J.C., Cooke, N.J.: Changes in team cognition after a retention interval: The benefits of mixing it up. J. Exp. Psychol. Appl. 17(4), 303–319 (2011)
11. Tognoli, E., Lagarde, J., DeGuzman, G.C., Kelso, J.A.S.: The phi complex as a neuromarker of human social coordination. Proc. Natl. Acad. Sci. U S A 104(19), 8190–8195 (2007)
12. Lindenberger, U., Li, S.C., Gruber, W., Müller, V.: Brains swinging in concert: cortical phase synchronization while playing guitar. BMC Neurosci. 10, 22 (2009)
13. Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., Garnero, L.: Inter-brain synchronization during social interaction. PLoS One 5(8), e12166 (2010)
14. Schippers, M.B., Roebroeck, A., Renken, R., Nanetti, L., Keysers, C.: Mapping the information flow from one brain to another during gestural communication. Proceedings of the National Academy of Sciences 107(20), 9388–9393 (2010)
15. Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T., Haynes, J.D.: Flow of affective information between communicating brains. NeuroImage 54(1), 439–446 (2011)
16. Astolfi, L., Toppi, J., De Vico Fallani, F., Vecchiato, G., Salinari, S., Mattia, D., Cincotti, F., Babiloni, F.: Neuroelectrical hyperscanning measures simultaneous brain activity in humans. Brain Topography 23, 243–256 (2010), doi:10.1007/s10548-010-0147-9
17. Dodel, S., Cohn, J., Mersmann, J., Luu, P., Forsythe, C., Jirsa, V.: Brain signatures of team performance. Foundations of Augmented Cognition. Directing the Future of Adaptive Systems, 288–297 (2011)
18. Stevens, R.H., Galloway, T.L., Wang, P., Berka, C.: Cognitive neurophysiologic synchronies: What can they contribute to the study of teamwork? Human Factors: The Journal of the Human Factors and Ergonomics Society (2011)
19. Tognoli, E., Kovacs, A., Suutari, B., Afergan, D., Coyne, J., Gibson, G., Stripling, R., Kelso, J.: Behavioral and brain dynamics of team coordination part i: Task design. Foundations of Augmented Cognition. Directing the Future of Adaptive Systems, 257–264 (2011)
20. Scholz, J., Schöner, G.: The uncontrolled manifold concept: identifying control variables for a functional task. Experimental Brain Research 126(3), 289–306 (1999)
21. Bressler, S.L., Tognoli, E.: Operational principles of neurocognitive networks. International Journal of Psychophysiology 60(2), 139–148 (2006)
22. Dumas, G., Chavez, M., Nadel, J., Martinerie, J.: Anatomical connectivity influences both intra- and inter-brain synchronizations. PLoS One 7(5), e36414 (2012)
23. Price, C.J., Friston, K.J.: Degeneracy and cognitive anatomy. Trends Cogn. Sci. 6(10), 416–421 (2002)
24. De Jaegher, H., Di Paolo, E., Gallagher, S.: Can social interaction constitute social cognition? Trends Cogn. Sci. 14(10), 441–447 (2010)

Analysis of Semantic Content and Its Relation to Team Neurophysiology during Submarine Crew Training

Jamie C. Gorman[1], Melanie J. Martin[2], Terri A. Dunbar[1],
Ronald H. Stevens[3], and Trysha Galloway[3]

[1] Texas Tech University, Lubbock, TX, USA
[2] California State University—Stanislaus, Stanislaus, CA, USA
[3] The Learning Chameleon Inc., Culver City, CA, USA
jamie.gorman@ttu.edu

Abstract. A multi-level framework for analyzing team cognition based on team communication content and team neurophysiology is described. The semantic content of team communication in submarine training crews is quantified using Latent Semantic Analysis (LSA), and their team neurophysiology is quantified using the previously described neurophysiologic synchrony method. In the current study, we validate the LSA communication metrics by demonstrating their sensitivity to variations in training segment and by showing that less experienced (novice) crews can be differentiated from more experienced crews based on the semantic relatedness of their communications. Cross-correlations between an LSA metric and a team neurophysiology metric are explored to examine fluctuations in the lead-lag relationship between team communication and team neurophysiology as a function of training segment and level of team experience. Finally, the implications of this research for team training and assessment are considered.

Keywords: Latent Semantic Analysis, Team cognition, Team communication, Team neurophysiology, Teamwork.

1 Introduction

A team is an interdependent group of two or more people who work together for a fixed amount of time to achieve a common goal [1-2]. Across a wide range of work environments, including business, military, medical, academic, and culinary settings, there are many common goals that are either too physically or cognitively demanding to be achieved by individuals working alone. To meet such goals, tasks must be performed in real time by people working together as a team. This paper focuses on the communicative and neurophysiological aspects of team cognition as crews work together to solve navigation problems and coordinate solutions in a submarine crew training environment.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 143–152, 2013.

1.1 Levels of Analysis in Team Cognition

Team cognition has been defined theoretically as either the shared declarative knowledge of team members, the *shared cognition perspective* [e.g., 3-5], or as the dynamic interactions (e.g., communications) between team members, the *interactive theory of team cognition* [6]. Inspired by the interactive theory of team cognition, we take a multi-leveled approach to studying team cognition as it unfolds across segments of submarine crew training. In this research, our overarching focus in on how team communication and team neurophysiology function as related metrics for team cognition, though each is concerned with a different level of analysis.

At a more "micro" level of analysis, we have analyzed neurophysiological patterns of team members as they work together to acquire team skill [7]. We argue that solely focusing on individual neurophysiology does not capture the dependencies that develop across team members as they acquire overt interaction patterns. The team neurophsysiology paradigm developed by Stevens and colleagues [7-8] addresses this by allowing one to examine distributions of neurophysiological patterns as they develop across team members. We have also developed a variety of approaches for studying team cognition at a more "macro" level by focusing on overt interaction patterns during team skill development [9]. Those overt measures of team cognition include behavioral [e.g., 10] and communication-based [e.g., 11] metrics. The major theme of this paper is to continue to extend our communication-based metrics of team cognition into a submarine crew training environment, while beginning to develop a framework to support a joint team communication—team neurophysiology paradigm for understanding team skill development.

The team neurophysiology methods considered in this paper are described elsewhere [7-8]. Therefore, we want to devote much of this paper to describing our analysis of team cognition during submarine crew training using semantic content analysis of team communication. Hence, we devote the following section to team communication analysis.

1.2 Analyzing Team Communication Using Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a mathematical/statistical method for representing and analyzing semantic knowledge within a domain [12]. LSA is based on the theory that knowledge is reflected in the contextual usage of words within meaningful discourse [13]. LSA takes as its input a raw corpus of text and represents the corpus as a matrix of unique terms (e.g., words) by documents (e.g., paragraphs). LSA assumes that lower-dimensional (latent) semantic factors account for the frequency of co-occurrence between words and documents in the raw matrix. The space of factors is called the semantic space, and it is constructed through singular value decomposition. The optimal number of dimensions can be determined such that the relationships between words and context results in correct inductions (e.g., synonym matching; missing word replacement).

LSA has been used to successfully distinguish high-performing from low-performing unmanned air vehicle (UAV) teams by comparing their transcribed communications to a UAV semantic space [11]. For the current study, we constructed a

semantic space representing nautical navigation knowledge with which to analyze submarine crew communication. The semantic space (314 factors, 124,326 total words, 6,846 terms, and 5,904 documents) was constructed from a corpus created from submarine crew training transcripts, the Navigational Rules of the Road (COLREGS), Submarine Operations Manual, and the unclassified Doctrine for Submarine Interior Communications.

1.3 The Current Study

The primary goals of the current study are to (a) validate semantic content metrics for identifying critical differences between submarine crew training segments, (b) use semantic relatedness metrics to differentiate between levels of submarine crew experience, and (c) demonstrate the dependent relationship between team semantic content and team neurophysiology in a submarine crew training environment. Finally, we consider the implications of our results for team training and assessment.

2 Method

2.1 Participants

The data used in the current study were collected from Junior Officer Navigation teams enrolled in the Submarine Officer Advanced Candidacy class at the US Navy Submarine School. These teams consisted of six or more crew members, including: Quartermaster on Watch; Navigator; Officer on Deck; Assistant Navigator; Contact Coordinator; and Radar. (Other team members were also present and participated, but were not analyzed using the neurophysiological methods described later.) These teams participated in Submarine Piloting and Navigation (SPAN) simulation sessions during the class. We analyzed seven of these SPAN sessions: Four are from more experienced teams, and three are from less experienced ("novice") teams. In the statistical analyses we present below, we use a between-subjects variable, Experience, to index Novice vs. Experienced SPAN training sessions.

2.2 Training Simulation

The communication metrics analyzed in the current study were calculated from transcripts of team communication across crews during SPAN training simulations. The SPAN sessions are high-fidelity training simulations that consist of three segments: Briefing; Scenario; and Debriefing [7-8]. During the Briefing segment, the overall goals of the mission are presented and discussed. The Scenario is the dynamically evolving segment of the training, during which teams navigate through a route in a high-fidelity submarine simulation. The Scenario segment requires teams to steer and change course or speed while identifying landmarks and other ships that factor into SPAN. During the Scenario, the team must also periodically take Rounds, during which three navigation points are chosen, and the bearing of each point from the boat

is measured and plotted on a chart. The accuracy and variability of Rounds may serve as a team performance measure in future research, but those data will not be analyzed here. The Debriefing segment is an after-action-review, during which teams discuss what worked and what other options or actions could have been taken during the Scenario. The Debriefing segment provides a teaching experience, where both short- and long-term learning goals are discussed. In the statistical analyses we present below, we use a within-subjects variable, Training Segment, to index the Briefing, Scenario, and Debriefing training segments. In addition to communication metrics, neurophysiology data taken from crew members during the training simulations are used in some analyses presented below.

2.3 Measures

Two metrics derived from the geometrical interpretation of the semantic space are (1) the *vector length* of a piece of discourse and (2) the *cosine* between two pieces of discourse. We calculated both of these metrics for seven SPAN transcripts (i.e., the four experienced team sessions and three novice team sessions). These metrics will be used to analyze the semantic content of team communication during SPAN.

The vector length of a piece of discourse (e.g., an utterance; "Recommend steering course 178 to regain track.") is the Euclidean norm of the vector, created by summing the semantic space vectors of words in the discourse, plotted in the semantic space. The vector length measures the amount of semantic content (cf. knowledge) a piece of discourse contains relative to the domain of discourse, as represented by our SPAN semantic space.

The cosine between any two pieces of discourse (e.g., any two utterances; any two training segments; any two complete transcripts; etc.) is the vector dot product between two vectors plotted in the semantic space. The correlation between two vectors can be shown to be the cosine of the angle joining them (e.g., independent, perpendicular vectors have $\cos[90^{o}] = 0$, and they are completely uncorrelated). Hence, the cosine measures the degree of semantic relatedness, or correlation, between any two pieces of discourse.

The team neurophysiological measure we will use (*NS Entropy*) to examine the relationship between communication content and team neurophysiology is derived from the EEG-based Neurophysiological Synchrony (NS) method, which is more fully described elsewhere [7-8]. Using this method, discrete, team-level NS states are sampled at a fixed interval (we used 1 Hz) from continuous EEG streams collected from each team member. The EEG-to-NS mapping is such that each discrete NS state identifies a different distribution of cognitive engagement (or workload; not analyzed here) across team members. As training segments unfold, the team engagement distribution changes, and is captured in a time series of discrete NS states. The set of NS states for SPAN was determined using an artificial neural network approach [7-8], which resulted in a set of 25 discrete NS states.

Though the cardinality of the NS states is fixed, there is no inherent numerical ordering of states. To quantify NS organization, therefore, we calculated the Shannon entropy across NS states using a sliding window of size 100s.

$$NS\ entropy = -\sum_{i=1}^{\#NS\ States} p_i \cdot \log p_i\ , \tag{1}$$

where p_i is the relative frequency of NS state i over a 100s window, was repeatedly calculated as the 100s window slid over the original, discretely-varying NS time series. Using this technique, for an input NS time series of length N, the output is a continuously-varying NS entropy time series of length $N - 99$. In this way, we use the first 100 samples to calculate the first entropy value at time $t = 100$, samples 2 through 101 to calculate the second entropy value at $t = 101$, and so forth. Using a window smaller than 100s has been found to increase the potential for false (discontinuous) spikes in the NS entropy time series [7]. In terms of team cognition, low entropy may be interpreted as a highly-ordered team neurophysiological state, whereas high entropy corresponds to a more random mix of team neurophysiological states [7].

3 Results

3.1 Differentiating between Task Phases Using Vector Length

To examine whether Training Segment and Experience underlie communication differences captured by LSA metrics, we first computed mean vector length across utterances for Briefing, Scenario, and Debriefing segments of each transcript and then we analyzed those mean vector lengths using a 3 (Training Segment) × 2 (Experience) mixed ANOVA. (We also analyzed cosines taken between successive utterances using this approach; however, none of those results were significant.) As illustrated by the vector length data shown in Figure 1, there was a significant main effect of Training Segment, F (2, 10) = 15.78, p = .001, η^2 = .76. No other omnibus effects were significant. A follow-up Tukey test on Training Segment (α_{FW} = .05) revealed that mean Debriefing vector length (M = 1.21; SD = .27) was significantly greater than both Briefing (M = .79; SD = .30) and Scenario (M = .51; SD = .03) mean vector lengths.

Careful reading of the utterances in the transcripts clearly indicated that teams were communicating differently as a function of Training Segment. Specifically, teams communicated with shorter, to-the-point utterances during the Scenario segment and longer, conversational utterances during the Briefing and Debriefing segments. To quantify this observation, we controlled for word count by dividing each utterance's vector length by the number of words in each utterance. The resulting quantity measures the rate of semantic content per word in each utterance: Communication Efficiency = Vector Length / Word Count [11]. To examine whether Training Segment and Experience underlie communication differences captured by the Communication Efficiency measure, we computed mean Efficiency across utterances for Briefing, Scenario, and Debriefing segments in each transcript and then analyzed those mean

Efficiency values using a 3 (Training Segment) × 2 (Experience) mixed ANOVA. The main effect of Training Segment was significant, F (2, 10) = 25.81, $p < .001$, $\eta^2 = .84$. No other omnibus effects were significant. A follow-up Tukey test on Training Segment ($\alpha_{FW} = .05$) revealed that Communications Efficiency was significantly higher during the Scenario ($M = .078$; $SD = .001$) than during the Briefing ($M = .054$; $SD = .006$) and Debriefing ($M = .039$; $SD = .003$) segments.

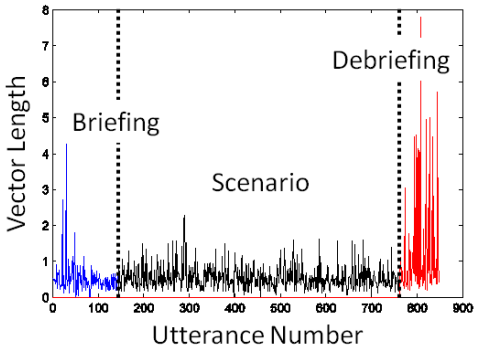


Fig. 1. Vector length of each utterance for an experienced team separated by Training Segment

3.2 Differentiating between Experienced and Novice Teams Using Cosines

To determine whether experienced teams' semantic content was more similar to each other than to novice teams and that novice teams' semantic content was more similar to each other than to experienced teams, we first calculated the LSA cosine metric between all possible pairs of transcripts as a function of Training Segment. We show the cosine matrix for the Scenario training segment in Table 1. If it is the case that semantic relatedness differentiates between experienced and novice team communication, then the bold values in Table 1 should be larger than the italicized values. We further examined that qualitative grouping using cluster analysis and multidimensional scaling (MDS).

A hierarchical cluster analysis of the Scenario cosine matrix (Table 1) using average between-groups linkage revealed that experienced teams clustered together and novice teams clustered together based on the semantic content of their communications for the Scenario training segment (Figure 2a). Similarly, a two-dimensional MDS solution for the Scenario cosine matrix (*Stress* = .86; $R^2 = .96$) revealed an "Experience" dimension, with novice teams low and experienced teams high on this dimension, and a second dimension that also appears to differentiate between teams by an as yet unidentified factor (Figure 2b). Hierarchical clustering and MDS conducted on the Briefing and Debriefing cosine matrices revealed that novice teams tended to be more tightly grouped in terms of semantic content than experienced teams during those training segments.

Table 1. Cosine Similarity Matrix Computed between All Pairs of Transcripts during Scenario

	Exper. 1	Exper. 2	Exper. 3	Exper. 4	Novice 1	Novice 2	Novice 3
Exper. 1	1.00						
Exper. 2	**0.91**	1.00					
Exper. 3	**0.81**	**0.85**	1.00				
Exper. 4	**0.85**	**0.87**	**0.82**	1.00			
Novice 1	*0.82*	*0.85*	*0.78*	*0.81*	1.00		
Novice 2	*0.76*	*0.77*	*0.72*	*0.77*	**0.88**	1.00	
Novice 3	*0.81*	*0.81*	*0.79*	*0.84*	**0.85**	**0.83**	1.00

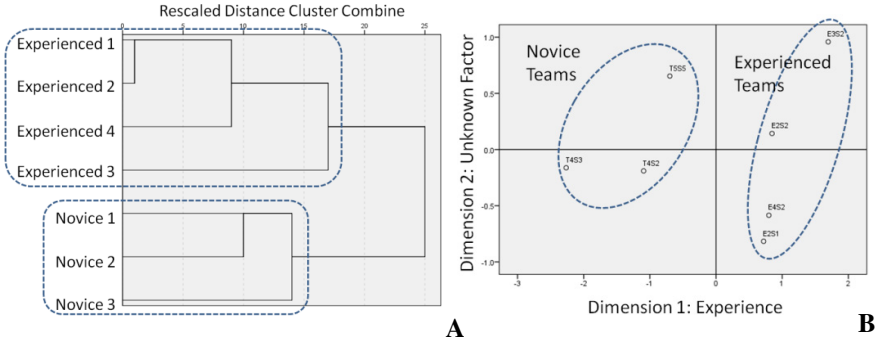


Fig. 2. (A) hierarchical clustering of the LSA cosine matrix from the Scenario training segment; (B) multidimensional scaling of the LSA cosine matrix for the Scenario training segment

3.3 Cross-Correlations between Semantic Content and Team Neurophysiology

Having established that (a) LSA-based vector length metrics differ as a function of training segment and (b) that LSA-based cosine metrics differentiate between experienced and novice teams, we turn to the question of how these differences in team communication are related to changes in team neurophysiology as a function of Training Segment and Experience.

We calculated the lagged cross-correlation function between LSA vector length of each utterance (Variable 1) and mean NS Entropy during each utterance (Variable 2) for each combination of Training Segment and Experience. The peak cross-correlation between these two variables (e.g., Figure 3) was identified to determine whether semantic content was leading (+ lag) or following (- lag) team neurophysiology and whether that correlation was significantly positive (+ direction) or negative (- direction). Table 2 provides basic information for each of the cross-correlations analyzed in the current study. (Because we were simply concerned with determining whether these variables were cross-correlated for the current study, we do not report or interpret exact values of lags and strength of correlation in this paper.) If the

correlation is significantly negative at a negative lag, as it is for the novice teams during the Briefing segment, then team neurophysiology is leading during that segment of training, such that higher entropy tends to temporally precede lower vector lengths. In terms of team cognition, this suggests that a more random mix team neurophysiological engagement states tends to temporally precede a reduction in the semantic content (cf. "knowledge") embodied in a team's communications, at least for some training segments and levels of team experience. Interestingly, we see the opposite cross-correlational pattern emerge for two of the three novice teams during the Debriefing segment. Though the cross-correlational patterns for the experienced teams are more varied, their interpretation can be carried out in the same way as our interpretation of novice teams' patterns.

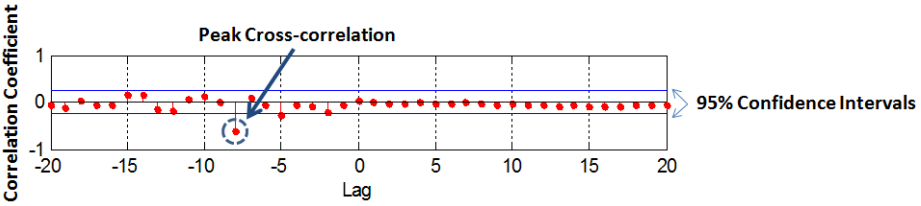


Fig. 2. Cross-correlation function for a novice team during the Briefing segment

4 Discussion

The LSA-based vector length metric of team communication content significantly differed as a function of training segment, and the cosine metric allowed us to differentiate between experienced and less experienced submarine crews. Those results lend support to our expectation that LSA-based metrics of semantic content can successfully distinguish teams of different skill levels and under different task constraints (e.g., planning a scenario vs. actually performing the scenario). The LSA communication efficiency measure also differed as a function of training segment. This metric is a hybrid between semantic content and simple syntactic markers of communication and represents just one of many possibilities for augmenting LSA metrics by adding contextual and syntactic details of team communication. In future research, each of these metrics should be validated against objective team performance measures (e.g., taking Rounds).

The lagged cross-correlations were largely significant and seem to suggest that team communication and team neurophysiology may lead or lag each other at different stages of team training or experience. Interestingly, we saw a mix of positive and negative peak cross-correlations, indicating that at times higher synchronization at the neurophysiological level is associated with increased domain specific semantic content at the verbal communication level and at other times with decreases in domain specific communication content. The prevalence of significant cross-correlation between team neurophysiogy and communication content begins to lend support to a joint team communication—team neurophysiology paradigm for understanding team skill development; however, the variety of lead-lag patterns and directions of cross-correlation must be disentangled in future research.

Table 2. Lag and Direction of Peak Cross Correlation between LSA Vector Length and NS Entropy as a Function of Training Segment and Experience

	Training Segment					
	Briefing		Scenario		Debriefing	
Experience	Lag	Direction	Lag	Direction	Lag	Direction
Exper. 1	-	+ *	-	+ *	+	- *
Exper. 2	+	+	+	+	+	+ *
Exper. 3	-	- *	-	+ *	-	+ *
Exper. 4	-	- *	-	+ *	+	+ *
Novice 1	-	- *	-	- *	+	+ *
Novice 2	-	- *	-	+ *	-	+ *
Novice 3	-	- *	-	- *	+	+ *

Note. * This correlation lies beyond the 95% Confidence Interval for no correlation; $p < .05$.

Finally, in this paper we have suggested a multi-leveled analysis of team cognition, which has implications for team training and assessment. During team development, the goal of team performance may be reflected in a variety of adjustments in the neurophysiological and overt behavioral patterns exhibited by teams as they learn to work together. In keeping with the interactive theory of team cognition [6], and similar to the theory of embodied cognition [14], neurophysiological patterns may constrain behavior patterns, or vice versa, at critical points during team skill development. As we are beginning to see with cross-correlation analyses, hierarchical patterns of constraint, in the form of lead-lag relationships between team neurophysiology and team communication, may be significantly altered by type of training and level of team experience. Although the exact nature of these developmental transitions remains to be seen; in the future, a joint neurophysiological/communication analysis may be critical for assessing key transitions in neural/cognitive team development.

Acknowledgments. This research is supported by Defense Advanced Research Projects Agency Contract W31P4Q-12-C-0166 and National Science Foundation Small Business Innovation Research Grant IIP 1215327. The findings, views, and opinions expressed in this paper are the authors' and do not necessarily represent the official views of any funding agency.

References

1. Cooke, N.J., Salas, E., Cannon-Bowers, J.A., Stout, R.: Measuring team knowledge. Human Factors 42, 151–173 (2000)
2. Salas, E., Dickinson, T.L., Converse, S.A., Tannenbaum, S.I.: Toward an understanding of team performance and training. In: Swezey, R.W., Salas, E. (eds.) Teams: Their Training and Performance, pp. 3–29. Ablex, Norwood (1992)

3. Cannon-Bowers, J.A., Salas, E., Converse, S.: Shared mental models in expert team decision making. In: Castellan, N.J. (ed.) Individual and Group Decision Making, pp. 221–246. Lawrence Erlbaum Associates, Hillsdale (1993)
4. DeChurch, L.A., Mesmer-Magnus, J.R.: The cognitive underpinnings of effective teamwork: A meta-analysis. Journal of Applied Psychology 95, 32–53 (2010)
5. Langan-Fox, J., Code, S., Langfield-Smith, K.: Team mental models: Techniques, methods, and analytic approaches. Human Factors 42, 242–271 (2000)
6. Cooke, N.J., Gorman, J.C., Myers, C.W., Duran, J.L.: Interactive team cognition. Cognitive Science. [Electronic publication ahead of print] (2012)
7. Stevens, R.H., Gorman, J.C., Amazeen, P., Likens, A., Galloway, T.: The organizational dynamics of teams. Nonlinear Dynamics, Psychology and Life Sciences 17, 67–86 (2013)
8. Stevens, R., Galloway, T., Wang, P., Berka, C.: Cognitive neurophysiologic synchronies: What can they contribute to the study of teamwork? Human Factors 54, 489–502 (2012)
9. Cooke, N.J., Gorman, J.C.: Interaction-based measures of cognitive systems. Journal of Cognitive Engineering and Decision Making 3, 27–46 (2009)
10. Gorman, J.C., Cooke, N.J., Amazeen, P.G., Fouse, S.: Measuring patterns in team interaction sequences using a discrete recurrence approach. Human Factors 54, 503–517 (2012)
11. Gorman, J.C., Foltz, P.W., Kiekel, P.A., Martin, M.J., Cooke, N.J.: Evaluation of latent-semantic analysis-based measures of team communications. In: Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting, pp. 424–428 (2003)
12. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284 (1998)
13. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240 (1997)
14. Shapiro, L.A.: Embodied cognition. Routledge, New York (2011)

Neurophysiological Predictors of Team Performance

Robin R. Johnson[1], Chris Berka[1], David Waldman[2], Pierre Balthazard[3], Nicola Pless[4], and Thomas Maak[4]

[1] Advanced Brain Monitoring, Inc., Carlsbad, CA, USA
[2] Arizona State University, W.P. Carey School of Business, USA
[3] St. Bonaventure University, School of Business, USA
[4] ESADE Business School, Ramon Llull University, Spain
{rjohnson,chris}@b-alert.com, waldman@asu.edu
pbalthazard@sbu.edu, {nicola.pless,thomas.maak}@esade.edu

Abstract. Objective: To identify benchmark neurophysiological measures that predict performance at a teaming level. Advanced Brain Monitoring has a track record of success in identifying neurophysiological metrics that impact expert behavior. For example, we characterized negative and positive predictors for marksmanship skill; persons with higher HF:LF Norm metrics of Heart rate variability (HRV, an indication of anxiety) during a benchmarking auditory passive vigilance task did not achieve expert marksman performance while those with above average visuospatial processing ability achieved greater levels of expertise. In the current research, we explored the ability of benchmark neurophysiological metrics to predict team performance in two large scale studies. Significance: Identifying neurophysiological metrics of teaming ability and performance as part of a team can provide potential screening mechanisms or developmental data to help build optimal teams and improve team interactions for different types of contexts in which teams may operate.

Keywords: leadership, neurophysiology, qEEG, prediction.

1 Introduction

Identifying leadership potential is a growing interest in military, academic, and industry applications. Many applications using personality profiling have been used to some success, however, these are self report mechanisms and do not access the internal processes that may contribute to leadership potential. Waldman and colleagues [1] argue that neurophysiology may provide more ecologically-valid assessment of psychological constructs associated with leadership.. Recent advances in the technical design of the qEEG hardware and software platforms enable practical application of qEEG in studying leadership potential during live teaming exercises. The main advantage of the qEEG-based team assessment is that it is continuous, and it does not require disruption of the ongoing team process. For example, [2] utilized the qEEG data for modeling team dynamics in complex military tasks.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 153–161, 2013.

Previous studies have found that neurophysiological profiles may identify predictors for those that were able to obtain expert status in a marksmanship task. The current study sought to determine the potential of using neurophysiological predictors for identifying leaders.

2 Materials and Methods

2.1 Participants

The students at a business school in the U.S. (Arizona State University) formed 43 teams of either 4 or 5 individuals. The overall sample comprised 201students with the mean age of 24.28 years. The participants were ethnically diverse (63.5% were Caucasian, 14.2% were Asian, and17.3% were Hispanic) and gender balanced (54.2% were males).

The students at a business school in Europe (ESADE, Barcelona) formed 31 teams of either 4 or 5 individuals. The overall sample comprised 146 students with the mean age of 28.7 years. The participants were ethnically diverse (61.5% were Caucasian, 20.7% were Asian, and 15.6% were Hispanic) and gender balanced (64.4% were males).

2.2 Protocol

All subjects were asked to complete a set of 3 benchmark tasks with simultaneous EEG: a 3-choice active vigilance task (3CVT), an auditory passive vigilance task (APVT), and a visual passive vigilance tasks (VPVT). Subjects completed these tasks as individuals before being assigned to teams of 3-5 people. The 3CVT required subjects to discriminate one primary target (presented 70% of the time) from two secondary non-target geometric shapes that were randomly interspersed over a 20 min period. Participants were instructed to respond as quickly as possible to each stimulus. A brief training period was provided prior to the start of the task to minimize the practice effects. The VPVT asked participants to keep pace with a visual stimuli that was presented every 2 seconds. Participants were instructed to depress the space key each time the stimuli was presented. The APVT was identical to the VPVT except that an auditory stimuli was presented in the APVT.

In the first study, we had advanced, business undergraduate students in a leadership course at Arizona State University attempt to solve the Ethical Decision Challenge™ from Human Synergistics/Center for Applied Research, Inc. of Chicago, Illinois. The exercise requires participants to rank 10 biomedical and behavioral research practices (all of which involve human subjects) in terms of their relative permissibility and acceptability [3-4]. This provides participants with an opportunity to engage in ethical analysis and decision-making. Examples of the practices include:

- Withholding study design information on purpose from participants when such information might skew their behavior within the study

– Conducting high risk but very important research when: a) there is no direct benefit to the participants, b) subjects are fully informed about the research and its risks, and c) they are capable of deciding whether or not to participate

While monitored for qEEG and qECG assessment, participants were given 5 minutes to initially read the problem statement, followed by an additional 10 minutes to provide their respective, individual solutions. They then repeated the task in a team process involving 4 or 5 individuals. The goal was to find a solution that all team members could "live with". The team process lasted 30-45 minutes. This allowed teams to complete the task without excessive time pressure and without generating participant fatigue or disinterest. To derive performance scores, solutions for both individuals and teams can be compared against expert scoring. In this study, the average solution of 800 members of Institutional Review Boards (IRBs) served as the normative scores. These IRBs had been established by various hospitals, universities, and research organizations throughout the U.S.

In the second study, conducted at the ESADE business school in Barcelona, Spain, the problem solving task addressed a corporate social responsibility case of the Levi Strauss Company involving child labor issues in Bangladesh [5]. Over approximately 40 minutes, students initially read the case (as individuals), formed a solution to the issues mentioned in the case, and recorded their respective solutions through a computer interface. After being fitted for qEEG and qECG assessment, they then engaged in a team discussion process involving 4 or 5 individuals. The goal was to derive a common solution to the issues mentioned in the case. The team process lasted up to an hour, including time for the recording (by one of the team members) of a solution onto a computer file. To derive performance scores, solutions for both individuals and teams were rated by two trained coders in terms of effective problem solving, decisiveness, and level of ethical development displayed in those solutions. The coders worked independently and showed high levels of inter-rater reliability in their scoring.

2.3 Leadership Performance Metrics

Both studies rated leadership similarly, although the second was more fine grained. The leadership scores involved other team member's assessment (for each respective team member) through a survey at the conclusion of the team task. In the first study, only shared leadership was assessed, while in the second, we added a more fine grained assessment.

In the second study, Leadership scores for each subject were assessed by the other team members in a survey that covered the following aspects of leadership:

- *transformational leadership* [6-7] - intellectual stimulation (i.e., helping others to examine and solve problems in new ways) and inspirational motivation (i.e., expressing confidence and enthusiasm about goals and what needed to be accomplished)

- *emergent leadership* [8-9] - the overall degree to which the team members relied on and considered a respective team member to have shown the leadership role during the team task.

All members of a respective team rated the other members (excluding himself). As the level of agreement among the subjects was high, these scores were averaged to provide a single score for each leadership measure for each subject. In the second study, these scores were averaged to identify the overall leadership score for each individual. The leadership scores were then ranked by team, with those with the highest scores categorized as "Leaders", those with the lowest scores assigned the category of "non-leader", and those in the middle assigned "Team-member". These categories were then used in the ANOVA to examine what neurophysiological metrcs are predictive of leadership role.

2.4 qEEG/ECG Data Recording and Signal Processing

The wireless B-Alert sensor headset [10] was used to acquire qEEG data of all subjects during the benchmark sessions. The qEEG recordings during the team process were synchronized with the respective videos. The qEEG data from 9 sites (POz, Fz, Cz, C3, C4, F3, F4, P3, and P4) were recorded with a sampling rate of 256 samples per second. The qEEG signals were first filtered with a band-pass filter (0.5-65Hz) before the analog to digital conversion and then the sharp notch filters were applied to remove environmental artifacts from the power network. The algorithm [11] was utilized to automatically detect and remove a number of artifacts in the time-domain qEEG signal, such as spikes caused by tapping or bumping of the sensors, amplifier saturation, or excursions that occur during the onset or recovery of saturations. Eye blinks and excessive muscle activity were identified and decontaminated by an algorithm [11] based on wavelet transformation.

From the filtered and decontaminated qEEG signal, the absolute and relative power spectral densities (PSD) were calculated on an epoch-by-epoch basis for each 1Hz bin from 1 to 40 Hz by applying fast Fourier transformation (FFT) to the 50% overlapping 1sec overlays of the qEEG data. In order to reduce the edge effect, the Kaiser window was applied to each overlay. Furthermore, the FFT on three successive overlays was averaged to decrease epoch-by-epoch variability. The following PSD bandwidths were extracted: theta slow (3-5 Hz), theta fast (5-7 Hz), theta total (3-7 Hz), alpha slow (8-10 Hz), alpha fast (10-12 Hz), alpha total (8-12 Hz), beta (13-30 Hz), and gamma (25-40 Hz).

In order to explore the applicability of neurological alertness quantification in estimation of the psychological metrics, we also included into the analysis the outputs of the B-Alert model [11-12] that quantifies engagement levels and identifies cognitive state changes. It is an individualized model that selects the most discriminative PSD variables, derives coefficients for a discriminant function, and

classifies subject's cognitive state for each epoch into one of the four levels of alertness: sleep onset, distraction/relaxed wakefulness, low engagement, and high engagement.

The p300 latency and amplitude components of the event related potential for the correct targets during the 3CVT task were also extracted. All individual trials of correct target responses were extracted; any trials that exceeded +/- 50 μV were removed, as were those with excessive artifact. All appropriate trials were then averaged and the maximum amplitude between 200-500 ms was determined, along with the latency of the maximal amplitude.

The B-Alert headset is enabled for the collection of heart rate, using a two lead set up, where one lead is place on the upper right collar bone and the other the lower left rib. Data is then sampled at 256Hz and the R-R spike indentified using proprietary algorithms, and the beat-to beat heart rate and heart rate variability measures are calculated per international standards [13].

2.5 Data Analysis

First, correlation analysis was performed to explore if any neurophysiological metrics were related to the leadership scores (shared in the first study, transformational and shared in the second).

As an initial investigation into what predictors might contribute to leadership development potential, we examined all neurophysiological metrics with 1-way ANOVA, comparing the leadership roles assigned (Leader, Non-Leader, Team-member) across the three benchmark tasks.

In order to explore the development of a predictive algorithm based on benchmark neurophysiology, we used the variables identified in the ANOVAs in a discriminate function analysis. We explored both a 3 class (Leader, Non-Leader, and Team Member), and a 2 class (Leader, Team Member) model.

3 Results

In this section the following results are presented: (1) statistically significant correlations between the neurophysiological measures during benchmark tasks and leadership at the individual levels, (2) ANOVA outcomes for neurophysiological metrics across leadership roles based on scores.

3.1 Correlations

Correlation analysis showed that small but significant correlations occurred between the individual level leadership scores and neurophysiological metrics. Table 1 shows the significant correlations for the first study (Shared leadership only), while table 2 shows the significant correlations for study 2 (transformational leadership, shared leadership).

3.2 ANOVA Results

No significant results occurred for the first study, indicating the single shared leadership metric is insufficient. For the second study ANOVA analysis revealed two distinctive patterns. First, during the VPVT, a passive visual vigilance task, we see significant increase in the Theta bands for the Leaders compared to the Non-Leader and Team-member categories. The ANOVA revealed a main effect for both Central Theta: $F(2, 134) = 3.29$, $p < .05$; and Left Theta: $F(2, 134) = 4.79$, $p < .01$. Post hoc analysis reveled that Leaders had the greatest activation, followed by the Non-Leaders, with the Team Members having the least activation. These data are shown in Figure 1.

A similar but inverse pattern of activation was found in the Frontal and Midline regions in Slow Theta (5-7 Hz), but not in overall or fast Theta during the Passive Auditory vigilance task (APVT). ANOVA revealed significant activation difference in the Slow Theta in the Frontal region: $F(2, 134) = 4.05$, $p < .05$, as well as the Midline region, $F(2, 134) = 4.05$, $p < .05$. Post hoc analysis found that while the Leadership role still had the greatest activation, the team members had the next greatest with the least being the Non-leaders.

Table 1. Correlations from Study 1compare subjectively scored shared leadership and neurophysiological metrics during an ethical decision making tasks associated with human subject informed consent protocols (Study1) and Child labor in developing countries (Study 2)

Neurophysiology Metrics	Pearson's R		
	Study 1	Study 2	
	Shared	Shared	Emergent
Frontal_AlphaSlow_8_10	0.206779		
Frontal_AlphaFast_10_13	0.236309		
Frontal_AlphaTotal_8_13	0.219375		
F3_Gamma_31_40			0.201425
Central_Gamma_31_40			0.200233
Left_Gamma_31_40			0.212858
HRV_pFreq_LFHFRatio		0.205656	
P300_Fz_Amplitude		0.371654	
P300_F3_Amplitude		0.40748	0.239428
P300_F4_Amplitude		0.216757	
P300_F4_Latency		-0.24146	
P300_C3_Amplitude	0.206123	0.279236	0.270103
P300_C4_Amplitude	0.263649		
P300_Cz_Latency	0.221524		
P300_POz_Amplitude	0.200265		
P300_P3_Amplitude	0.274717		
P300_POz_Latency		-0.32695	-0.34109
P300_P3_Latency	0.234837		
P300_POz_Amplitude	0.200265		

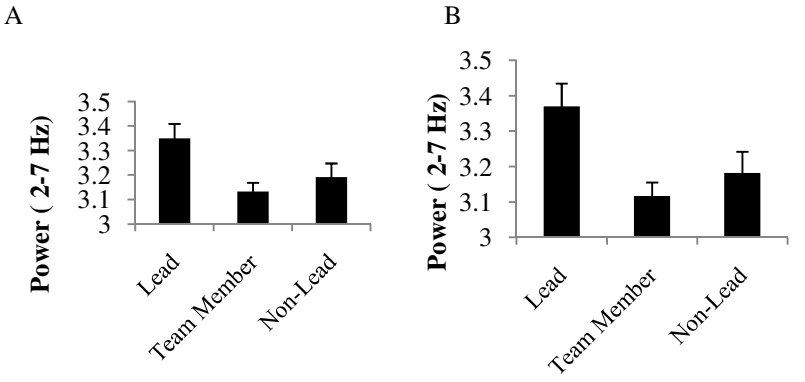


Fig. 1. Theta power (2-7 Hz) in the central (A) and left (B) scalp regions is significantly elevated during the Visual vigilance task for those that emerge as leaders during the ethical decision making task associated with child labor in developing countries. (* indicate significant post-hoc differences from Leaders, Ŧ indicate significant differences from Team Members).

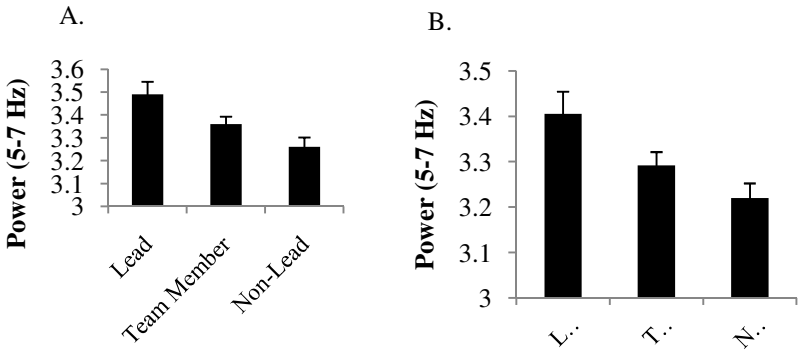


Fig. 2. Slow Theta power (5-7 Hz) in the frontal (A) and midline (B) scalp regions is significantly elevated during the Auditory vigilance task for those that emerge as leaders during the ethical decision making task associated with child labor in developing countries. (* indicate significant post-hoc differences from Leaders, Ŧ indicate significant differences from Team Members).

4 Discussion

The data shown herein demonstrate the potential of neurophysiological predictors of leadership development. Small but significant correlations were shown, particularly in the latency and amplitude of the P300 component across scalp sites, during the 3CVT. Longer latency P300s were associated with lower leadership scores. This finding is consistent with a body of literature linking latency slowing with decreased cognitive ability and slower response times in a variety of tasks [14-17]. ANOVAs revealed that central and left theta during the VPVT and frontal and midline theta during the APVT

predict the leadership role later taken during the ethical decision making task associated with child labor in developing countries. There were no such findings in the first study, with the human subject consent issues. This may be due to the inadequacy of the single metric of leadership.

The leadership status was based on team members rating each other, a highly subjective, but ecologically valid assessment. In addition, there is a high degree of variability of these scores. All teams had a clear leader, but the strength of that leader is variable, with the scale of the metric going from 1-5. Most leaders among the groups were ranked in the low to mid 4 range. However some were in the low 3 range. This may indicate low team coherence, dissonance in decision making, etc. In other teams we had several persons score above 4 (although one was always higher), perhaps indicating a strong group decision making process. Further breaking down the analysis by the strength of the leader may lead to additional finer grained analysis that may prove more helpful in identifying effective leaders, not just those most likely to emerge as a leader. In the second study, additional objective third party experts also ranked the leadership status of the team members. These scores may prove more informative than the internal, subjective measures taken and compared in the current analysis.

The tasks used herein were ethical decision making tasks. The "correctness" of each solution was also scored, in the second study, by expert judges. Allowing these metrics to be entered into the model of leadership may prove useful in future analysis.

References

1. Waldman, D.A., Balthazard, P.A., Peterson, S.J.: Social Cognitive Neuroscience and Leadership. The Leadership Quarterly 22(6), 1092–1106 (2011)
2. Stevens, R., et al.: Modeling the Neurodynamic Complexity of Submarine Navigation Teams. Computational and Mathematical Organization Theory (2012)
3. Balthazard, P.: Virtual version ethical decision challenge by R. A. Cooke. Human Synergistics/Center for Applied Research, Arlington Heights (2000)
4. Cooke, R.A.: The ethical decision challenge. Human Synergistics/Center for Applied Research, Arlington Heights (1994)
5. Pless, N., Maak, T.: Addressing Child Labour in Bangladesh. In: Stahl, G.K., Menderhall, M.E., Oddou, G.R. (eds.) Readings and Cases in International Human Resource Management and Organizational Behavior, 5th edn. Routledge, London (2011)
6. Balthazard, P.A., Waldman, D.A., Warren, J.E.: Predictors of the Emergence of Transformational Leadership in Virtual Decision Teams. The Leadership Quarterly 20(5), 651–663 (2009)
7. Bass, B.M., Avolio, B.J.: The Multifactor Leadership Questionnaire. Consulting Psychologists Press, Palo Alto (1990)
8. Carson, J.B., Tesluk, P.E., Marrone, J.A.: Shared Leadership in Teams: An Investigation of Antecedent Conditions and Performance. Academy of Management Journal 50(5), 1217–1234 (2007)
9. Zhang, Z., Waldman, D.A., Wang, Z.: A Multilevel Investigation of Leader-Member Exchange, Informal Leader Emergence, and Individual and Team Performance. Personnel Psychology 65(1), 49–78 (2012)

10. Berka, C., et al.: Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset. International Journal of Human-Computer Interaction 17(2), 151–170 (2004)
11. Berka, C., et al.: EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. Aviation Space and Environmental Medicine 78(5), B231–B244 (2007)
12. Johnson, R.R., et al.: Drowsiness/Alertness Algorithm Development and Validation Using Synchronized EEG and Cognitive Performance to Individualize a Generalized Model. Biologizal Psychology 87(2), 241–250 (2011)
13. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Circulation 93(5), 1043–1065 (1996)
14. Donchin, E., et al.: Cognition and event-related potentials. II. The orienting reflex and P300. Annals of the New York Academy of Sciences 425, 39–57 (1984)
15. van Nunen, P.E., Declerck, A.C.: P300, alertness and cognition. Acta Psychiatrica Belgica 94(2), 96–97 (1994)
16. Polich, J., Kok, A.: Cognitive and biological determinants of P300: an integrative review. Biological Psychology 41(2), 103–146 (1995)
17. Polich, J., et al.: P300 latency reflects the degree of cognitive decline in dementing illness. Electroencephalography and Clinical Neurophysiology 63(2), 138–144 (1986)

How Long Is the Coastline of Teamwork?

A Neurodynamic Model for Group and Team Operation and Evolution

John Kolm[1], Ronald Stevens[2], and Trysha Galloway[2]

[1] Team Results, USA
[2] UCLA School of Medicine, The Learning Chameleon, Inc., USA
trjohn.kolm@teamresultsusa.com, immexr@gmail.com,
ysha@teamneurodynamics.com

Abstract. A five-state Markov model is proposed for group and team operation and evolution that has a stronger basis in neurodynamics, greater descriptive accuracy and higher predictive value than many existing models. The derivation of this model from the symbolic analysis of normalized EEG activity during assigned team and group tasks is discussed, as are observations on team and group dynamics which emerge from the model. The predictive value of the model is shown when applied to independent data from submarine crew evolutions. Observations are offered on team dynamics which show the five-state model and its accompanying state transitions to be necessary and sufficient to describe both linear and non-linear team dynamics, and to begin unifying these traditional and new approaches in a straightforward way.

Keywords: nonlinear dynamics, neurodynamics teamwork, markov model, state transition, EEG symbol, tuckman.

1 Introduction

Ever since Benoit Mandelbrot [1] observed in his 1967 paper *How Long is the Coast of Britain?* that the apparent structure of complex dynamical phenomena can depend on the scale of magnification used, students of group and team dynamics have struggled to find the right observational lens through which the linear and non-linear dynamics of teams and organizations can be understood equally well. With a large observational aperture, gestalt states applying to a whole work team – for example the "forming, storming, norming, performing, adjourning" states well-known from Tuckman [2] – make excellent sense and are well understood. At small apertures drilling down toward individuals, non-linear states and behaviors where there is important fine structure and no useful gestalt characterization make equally good sense and are partly understood, albeit much less predictably in outcome.

The difficulty to date has been that team and group dynamics from a standpoint of workplace productivity are often best viewed through apertures of medium size. At these

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 162–171, 2013.

apertures, traditionally-understood linear phenomena and more-recently investigated non-linear phenomena become equally important and can each be crucial in determining practical productivity outcomes. A straightforward model of team and group dynamics and evolution which unifies both linear and non-linear phenomena is therefore an essential tool for the modern practical leader.

It is also desirable to base any such model, where possible, on verifiable and observable facts about human cognition. Inference drawn from behavior is important and remains the basis of much of psychology, but where it is possible to observe cognitive truth directly and thus to improve both the quality of observations and the insightfulness of descriptive models, opportunity exists for better science. Previously Stevens and colleagues [3] have taken advantage of technological developments in neuroscience and EEG monitoring to describe the neurodynamics of teams. This study extends these descriptions through the development of a five-state Markov model of teamwork which coalesces the complex phenomena into a simpler taxonomy that is well suited to practical team dynamics in industry and government.

2 Methods

2.1 Task and Teams

The custom-designed group task involved the team-based steering of a radio-controlled vehicle over an obstacle course and has been used for large-scale team training since 1996. The intention of the exercise was to present subjects with a significantly non-linear task to manage that kept the team motivated and engaged.

The operating area consisted of a Subject Zone within which subjects and experimenters were seated, and a Chicane Zone within which a radio-controlled vehicle, a varying number of chicanes, and four targets were located (Fig. 1). The Subject Zone contained seating for four subjects, each within easy reach of an individual controller for the vehicle steering system. Also within this area were a radio control system for the vehicle, a radiotelemetry monitoring and recording center for the subjects' EEG units, and a video camera to record video and audio. The Chicane Zone contained a small radio-controlled vehicle, four clearly-marked targets for the vehicle to strike, and a varying number and placement of wooden chicanes which was adjusted between the first and second task evolution. The targets each contained a detection system which caused them to emit clear visible and auditory feedback when struck and "set off" by the radio-controlled vehicle. Subjects were instructed that the goal was to use the vehicle to strike and "set off" all four targets in any order.

The radio-controlled vehicle operated like a tracked vehicle with steerage by wheels only. Ordinarily it would be a simple matter for a single operator to control this vehicle with a single radio remote, but the remote was replaced with a custom-built system which required four subjects to issue finely-coordinated commands in order to control the vehicle. Each subject was provided with a controller unit offering four buttons - left forward, right forward, left reverse and right reverse. Each function

would only operate the vehicle as commanded if all four subjects pressed the relevant button at the same time; moreover each function would not cease to operate until all four subjects released the relevant button at the same time.

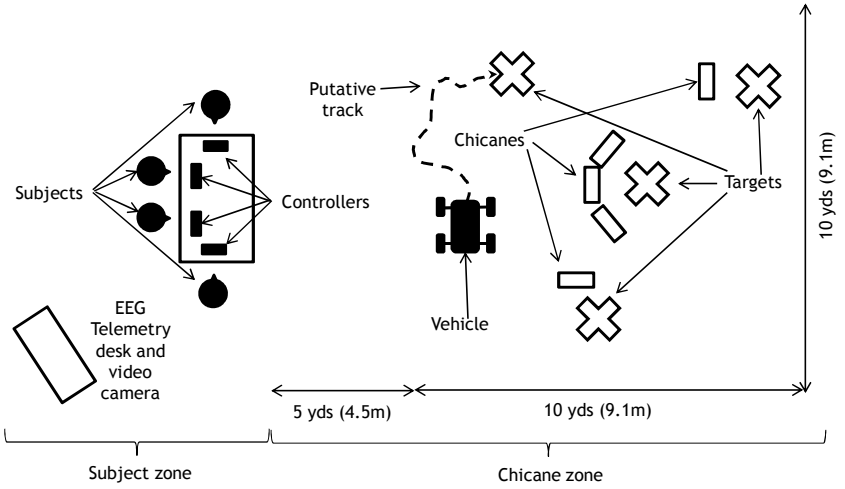


Fig. 1. Experimental layout for both task evolutions (chicane layouts vary)

The net effect of this system was twofold. Firstly, a high and constant level of engagement was required by the tight communication and feedback-management constraints of the task. Secondly, even with excellent team operation, a combination of mechanical tolerances, reaction time differences and uneven ground in the Chicane Zone meant that simple linear plans - such as the vehicle traveling in a straight line when appropriately commanded by the team – only worked for periods of a second or two. The combined effect yielded a non-trivial task in which linear and non-linear elements were combined in a way that could not be practically deconstructed.

Two task evolutions were performed that were characterized as “easy” and “hard”, with the “easy” evolution run first. In the “easy” evolution, only one or two chicanes were used, and none were placed in particularly awkward places with respect to the targets. In the “hard” evolution, targets were placed in more challenging locations within the Chicane Zone and more chicanes were used, some with awkward placing.

Subjects were also instructed to appoint a leader, and leadership was rotated after each target was “set off”, resulting in each subject being designated as the leader once per evolution, always in the same order.

The four subjects were tertiary-educated adults employed in the workforce by a range of employers, and not normally working together as a team. The same subjects were used for each task evolution, located in the same four physical positions, and with subject order preserved in the symbol elements generated for both. Subjects (n=4) performed the two task evolutions with a break in between. During each evolution of the task, all four subjects were simultaneously monitored by EEG.

2.2 Electroencephalography (EEG)

The B-Alert® system by Advanced Brain Monitoring, Inc. is an easily-applied wireless EEG system that includes software that identifies and eliminates multiple sources of biological and environmental contamination and allows second – by -second classification of cognitive state changes [3]. The 9-channel wireless headset includes sensor site locations: F3, F4, C3, C4, P3, P4, Fz, Cz, POz in a monopolar configuration referenced to linked mastoids. B-Alert® software acquires the data and quantifies engagement (EEG-E) in real-time.

For each task the four team members were rank ordered (4 = highest, 1 = lowest) with regard to the levels of EEG-E. The positions of the leaders in each performance were then compared with the average positions of the remaining team members. In all eight performances the leader had the highest or second highest levels of EEG-E (mean ranking Leaders = 3.34, Other Members = 2.21, $T = 4.80$, $df = 7$, $p < 0.002$).

3 Design and Procedure

3.1 Team Neurodynamics

For neurodynamics modeling, normalized second-by-second values of EEG-E were concatenated into vectors representing the levels being expressed by each team member. For instance, in Fig.2A team members 3 and 4 were expressing below average levels of EEG-E and would be assigned values of -1. Team members 1 and 2 were expressing above average levels of EEG-E and were assigned the value 3. A team member with average levels would be assigned the value 1; the vector representation was therefore (3, 3,-1,-1). Using unsupervised artificial neural networks (ANN) where the nodes were arranged in a linear configuration, the vectors from all performances were modeled into collective team variables that are termed neurodynamic symbols of engagement (NS_E). ANN classification of these second-by-second vectors created a symbolic state space showing the possible combinations of either EEG-E or EEG-WL across team members (Fig. 2A). One effect of the linear configuration of neural network nodes during ANN training is that symbols that resemble each other become closely aligned. For instance, in Fig. 2B NS 1-5 represented periods where most team members had average / below average levels of EEG-E while NS 20-25 represented times when most had above average EEG-E levels.

While a symbolic view of the state of the team is useful for characterizing team neurodynamics, it is not the best representation for quantifying team neurodynamics. Although there are methods for the quantitative representation of symbols, we chose a moving average window approach to derive numeric estimates of the Shannon entropy of the NS symbol stream [3]. Entropy is expressed in terms of bits; the maximum entropy for 25 randomly-distributed NS symbols would be log2 (25) or 4.64.

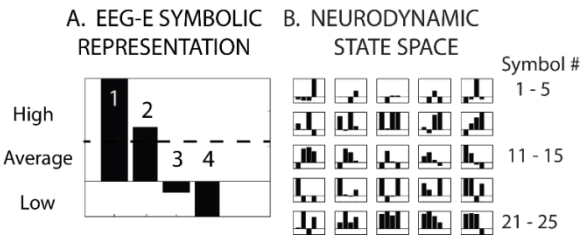


Fig. 2. Data Flow for Creating Team Neurodynamics Models. ANN classification of second-by-second vectors (A) creates a symbolic state space showing the possible combinations of EEG-E or EEG-WL across (numbered) members of the team (B).

For comparison, an entropy value of 3.60 would result if roughly half (12) of the NS symbols were randomly expressed. To develop an entropy profile over a session, the NS Shannon entropy was calculated at each epoch using a sliding window of the values from the prior 60 seconds. As teams entered and exited periods of organization, the entropy should fluctuate as a function of the number of NS symbols being expressed by the team during a block of time [3]. As shown in Fig. 3 for the hard problems, there were significant entropy fluxes, with the periods of greatest team organization (i.e. the lowest NS_E entropy) occurring around periods where there was a target hit, or an expected target hit.

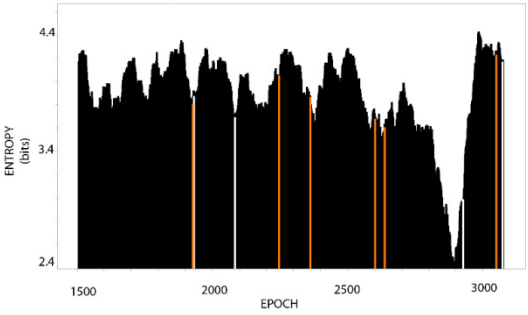


Fig. 3. NS_E Entropy Fluctuations. The fluctuations in the NS_E entropy levels are shown for the hard problems. The lines mark where there was a hit, or a near miss.

3.2 Symbols and Phase Transitions

One way of visualizing the short-term structural dynamics of a data stream is to create transition maps that plot symbol being expressed a time t vs. that at time $t + 1$; such maps are shown for the "easy" case (Fig. 4A) and "hard" case (Fig. 4B). An examination of the phase transition diagrams for the "easy" and "hard" cases reveals attractive basins along the diagonal in both cases, representing relatively stable symbols, and also off-diagonal attractors which indicate common symbol transitions. The hard problems showed fewer of the off-axis transitions indicating a more organized cognitive state. Randomizing the NS data stream destroyed this organization (Fig. 4C). As expected from the transition matrices, the harder tasks had lower overall NS_E

entropy levels. These transitions show a practical landscape of team preferences in the context of the task environment. In the "easy" case, symbols of particular interest on the diagonal are 5, 7 and 25. High-usage off-diagonal symbol transitions include 15-to-25, 25-to-15, 23-to-7 and 27-to 7. It is also apparent that while some symbol transitions are bilaterally symmetrical, for example 15/25, not all are. In the "hard" case, symbols of interest include 1, possibly 11 and 19, and 21 along the diagonal; and the off-diagonal transition 1-to-21. Similar observations about possible bilateral asymmetry apply. The phase colorings also denote considerable additional structure showing relationships of interest between symbols, but these seem numerous, complex, and confusing as they stand.

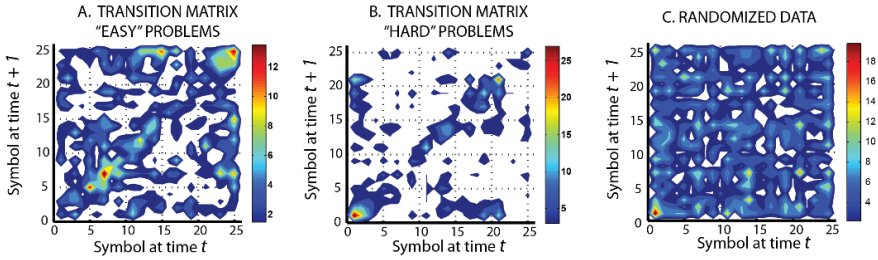


Fig. 4. Neurodynamic Symbol Transition Matrices for Easy (A) and Hard (B) Problems. Randomization of the combined data destroyed the structure (C).

The dimensionalities of the data streams were estimated by the Hurst exponent, where an exponent of 0.5 indicates a random process while an exponent between 0.5 and 1 indicates a persistent process, i.e. an upward or downward trend is likely to continue. The Hurst exponents for the data stream in Figures 4A and 4B were 0.88 and 0.67 respectively suggesting the NS data streams for these tasks have a fractal structure; i.e. a process somewhere between deterministic and random. As expected, randomizing the data stream reduced the Hurst exponent to 0.47.

4 Analysis

4.1 Development of a Symbol Taxonomy for Transitions

To extend the transition matrix representation, a taxonomy was applied to the major transitions in Fig. 4, focusing on the symbol transitions that were most heavily-used by the team while accomplishing both the "easy" and the "hard" tasks. The goal was to develop a taxonomy based on the distributions of EEG-E by different members of the team (Table 1); the motivation for this scheme was based on general principles from leadership development discussed later. The move from 25 EEG symbols to five underlying and descriptive and characterized states for the team – using the term "states" in the Markovian sense – is key, and the five Markov states (Dominant, Dyadic, Collegiate, Outlier, Dormant) are used subsequently. The outstanding questions, covered next, are how we can maximize the information yield of the data under this model, and whether the five states are necessary and sufficient.

Table 1. Taxonomy based on the distributions of EEG-E

Evolution	Symbol	Name	Characterization
Easy	5	Dominant	One person with high engagement; the rest follow with uniformly lower engagement.
Easy	7	Dyadic	One small clique with high engagement; the rest follow with uniformly lower engagement.
Easy	15	Outlier	One small clique with distinctively low engagement; the rest with much higher engagement.
Easy	23	Collegiate	Uniformly high and approximately equal engagement
Easy	25	Outlier	Ibid
Hard	1	Dominant	Ibid
Hard	11	Outlier	Ibid
Hard	21	Outlier	Ibid
All	2,3,4	Dormant	Uniformly low and approximately equal engagement

4.2 Data Aggregation

In both the easy and hard cases, 25x25 transition frequency matrices – the numerical, and accurate, counterpart of a colored phase transition diagram – were generated. Each symbol was assigned to a state in the taxonomy, and then the frequency transition counts for each state were aggregated. The resulting state transition tables were:

Table 2. Aggregated transition counts,"easy" case, row-column order

	COL	DOM	DOR	DYA	OUT
COL	57	34	9	46	60
DOM	22	91	31	85	94
DOR	11	44	28	29	31
DYA	53	80	41	113	102
OUT	63	74	33	117	155

Table 3. Aggregated transition counts, "hard" case, row-column order

	COL	DOM	DOR	DYA	OUT
COL	23	18	2	30	28
DOM	20	127	66	118	99
DOR	7	81	60	39	44
DYA	24	123	53	140	86
OUT	27	81	50	99	125

Aggregating counts in this way allows us to use the theoretical maximum information rate from the available data. Moreover, as we apply the taxonomy in part 4.1 to all 25 symbols, we observe that this taxonomy is necessary and sufficient to cover all symbols. There are no symbols that do not "fit", but if any one of the five states is removed from the taxonomy, this ceases to be the case.

4.3 The Markov Model

A Markov model offers the advantages of simplicity, practical and immediate usability by workplace managers, and a well-developed body of knowledge and understanding derived from uses in math, engineering and other areas of the life sciences [4]. Such a model posits a number of underlying states of a system and a collection of probabilities of transition from any state to any other, including itself. We can now take the state transition counts, convert these to probabilities and then map them into the following model for group and team operation and evolution (Fig. 5 and 6).

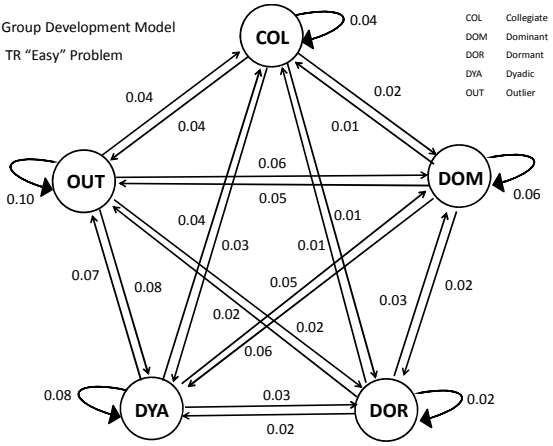


Fig. 5. Model for the "easy" case

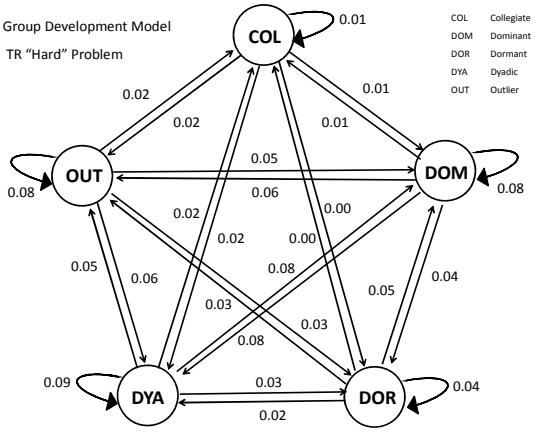


Fig. 6. Model for the "hard" case

Some preliminary observations about the dynamics of the subject team can be made from these models:

1. In the "easy" task, the team almost never transits from Collegiate to Dormant or Dormant to Collegiate. Therefore and if for example it should be undesirable for a team ever to be Dormant, with this dataset the safest state to try to engineer in a team would be Collegiate.
2. In the "hard" task, the bidirectional low-probability transition between Collegiate and Dormant is further accentuated, and simply never occurs. Thus this latent "forbidden" transition seems to be fundamental and is accentuated as the job becomes more demanding.
3. Although the team was forcibly started in the Dominant state by being instructed to select a leader, it does not remain in this state. In both the "easy" and "hard" tasks, Dyadic is slightly more stable state than Dominant. In the "hard" task, the hierarchical relationship between Dyadic and Dominant states is preserved, but Dormancy increases (perhaps owing to being "stumped" more often), Outlier behaviors reduce somewhat and Collegiate behavior drops significantly.
4. Outlier and Dyadic have a close relationship in the "easy" task, as do Dyadic and Dominant in the "hard" task. With clique leadership, minorities are often lost, dropping their engagement; and one overall leader still emerges frequently.

Now that we know the model is necessary and sufficient, and that it explains the observations, the remaining question is whether it has predictive value.

5 Discussion

The proposed taxonomy is also satisfying and robust from the standpoint of some 20 years' experience with team development in industry. We are all familiar with the "Dominant", one-leader team dynamic for example, as we are with leadership cliques in a "Dyadic" state; wholly engaged and disengaged teams in the Collegiate and Dormant states; and breakaway or disaffected minorities in the "Outlier" state. Examples of high-probability state transitions familiar to the experienced leader include the tendency of leadership cliques to disaffect some team members who feel ignored, and the rarity of truly collegiate and leaderless behavior. The model also applies well to an earlier study done on a large automotive company in which a rapid transition from Dominant to Outlier dynamics, which then became recursive, closed a manufacturing plant for two days at a cost of around $5M [4].

The model can also be shown to have predictive value, in that the same model was then applied to data from a previous experiment with submarine crews [3] and found to fit. Many of the same properties of team dynamics were verified with this data, and some new features applicable to submarine crews were also found.

We now also see the fusion offered by this new taxonomy between linear and nonlinear dynamics. This is especially useful in the mid-range organizational scale and

apertures of observation favored by workplace managers in which the effects of linear, gestalt emergent behaviors and non-linear, non-gestalt behaviors in work teams become equally important.

The science and the brain itself are telling us that we need to model two gestalt states and three non-gestalt states. The gestalt, whole-of-team states are Collegiate and Dormant; in these states, the team can indeed be lumped together and considered as one, as older models assume for all states. The Dominant, Dyadic and Outlier states however are non-gestalt states, in which the granularity and fine structure of the team must be taken into account. The modern manager can simply use the five states as a model, confident that both linear and non-linear events are catered for.

The new model has immediate application in government and industry for practical managers at the line, middle, senior and top levels. Previous models of team evolution often do not match well with real-world observation, and are synthetic rather than analytic. An analytic model permits diagnosis and correction. Managers can readily spot whether a team is in a Dominant, Dyadic, Collegiate, Outlier or Dormant state and can be fairly confident of the likely futures, allowing good decisions to be made quickly. Passive observation of synthetic models offers no such call to action.

The model also lends itself well to recruitment, team-based interventions that have a measurable effect on productivity, change management and – perhaps most importantly – to the promotion of good and simplifying science in industry.

Acknowledgements. Special thanks to Cory Lutz, Leslie Gruis, Amanda Biller, Lewis Carroll, Sanjay Mishra and Anne Wilburn in particular for their invaluable assistance. This work was supported in part by NSF SBIR grants IIP 0822020 and IIP 1215327.

References

1. Mandelbrot, B.B.: How long is the coast of Britain? Science 156, 636–638 (1967)
2. Tuckman, B.E.: Stages of Small Group Development. Group and Organizational Studies 2, 419–427 (1977)
3. Stevens, R., Galloway, T., Wang, P., Berka, C., Tan, V., Wohlgemuth, T., Lamb, J., Buckles, R.: Modeling the Neurodynamic Complexity of Submarine Navigation Teams. Comput. Math. Organ Theory (2012), doi:10.1007/s10588-012-9135-9
4. Kolm: Extinction at the Corporate K-T Boundary. In: Proceedings of the Society for Chaos Theiry in Psychology and the Life Sciences. 23rd Annual Conference. SCPTLS, Baltimore (2012)

Effects of Teamwork versus Group Work on Signal Detection in Cyber Defense Teams

Prashanth Rajivan[1], Michael Champion[1], Nancy J. Cooke[1], Shree Jariwala[1], Genevieve Dube[2], and Verica Buchanan[1]

[1] Arizona State University, CERTT Lab, 7418 S Innovation Way, Mesa, Arizona, 85212 USA
[2] Université Laval, 2325 Rue de l'Université, Quebec, QC G1V 0A6, Canada
pnrajiva@asu.edu

Abstract. Cyber security is critical for any modern day organization's operations. Organizational structure and reward policies not conducive for teamwork may be affecting the performance of cyber defense analysts. Past research shows that team interaction could lead to better cyber defense performance. However, the value of team work in the cyber defense context has not been demonstrated using empirical methodologies. To explore this, we conducted a study on the effects of teamwork versus group work (i.e., looking at both the team and individual levels) on signal detection performance of cyber security defense analysts using the synthetic task environment called CyberCog. The results from the preliminary analysis conducted reveal that simply encouraging analysts to work as a team and providing team-level rewards leads to better team performance in cyber defense analysis.

Keywords: Cyber Defense Performance, Team Cognition, Team Performance, Synthetic Task Environment, CyberCog.

1 Introduction

A recent study sponsored by Hewlett Packard [1] has found a 42% increase in the number of cyber attacks on organizations in the U.S. alone in 2012. To counter this growing number of cyber attacks, effective cyber defense is essential. Cyber security defense involves protecting the computer networks of an organization from any malicious activities such as malware attack and cyber espionage [2]. Personnel defending an organization's computer networks from cyber based attacks are often called cyber security defense analysts (or CSD analysts).

Cyber Defense Analysis. A typical organization contains a large number of computing systems such as desktop computers, laptops, servers, networking devices, and more that produce large amounts of data in the form of system logs, network traffic data and sensor data (alerts from intrusion detection systems (IDS)). CSD analysts have to monitor and fuse large amounts of data in order to identify patterns that may correspond to potential cyber attacks [3][4]. For example, analysts usually start from a

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 172–180, 2013.

suspicious set of intrusion alerts, filter network level data pertinent for those intrusion alerts, find associated system level logs, find intelligence reports relating to the situation, and then using their experience and training analyze the data collected to decipher if their network is being attacked or not. Once the analysts suspect there is an ongoing attack, the analysts start collecting data as evidence to support their suspicion and to eventually report the findings to higher authorities [2]. Finally, the analyst must assess the adversaries' intentions and capabilities to take the appropriate response. These tasks are mostly conducted manually using command level interfaces or graphical interfaces.

Due to cyber attacks evolving at very high speeds [5], this reduces the time available to respond to an attack, and adds to analyst's cognitive overload [6]. The cyber defense analysis task involves uncertainty with high information load and requires experienced personnel with domain knowledge. Because of this, CSD analysts are often placed under extreme time pressure. In some settings they have to process the alerts given to them at a pace of one every two minutes. Leading to further frustrations, alerts generated from current IDSs are often false alarms and thus the onus is on the analysts to distinguish the alerts that correspond to an attack from false alarms. Thus, a combination of factors that include overwhelming amounts of data, numerous false alarms, and time stress leads to cognitive overload in cyber defense analysts [6].

With cyber defense analysis being a complex task, it is sometimes performed by CSD analysts as a large group, with each analyst working on different levels of the task with specific domain knowledge and experience. However, simply bringing a group of people together to work on a task would not suffice. To work on such complex tasks we need actual teams of CSD analysts. What often occurs with CSD analyst teams is a loose association among individuals, rather than a functioning team [6]. For our definition, a team is a type of a group in which members of the team have diverse backgrounds, but work together in an interdependent manner towards a common goal [7]. Team cognition, which is defined as cognitive processes such as decision-making and learning, occurs at the team level [8] and has a significant effect on team performance [9][10]. Cooke and colleagues [9] proposed a theory of Interactive Team Cognition (ITC) which is a recent perspective on team cognition which states that team cognition emerges from team interactions. This is contrasting to the earlier theory of shared team cognition [10] which states that team cognition is the sum of the knowledge of individual team members. ITC does not however dispute the importance of individual knowledge for effective performance, but argues instead that team cognition is not solely tied to the knowledge of the individual members of the team.

As aforementioned, CSD analyst groups have been observed to lack teamwork [6]. We hypothesize that existing organizational structure and reward policies could be one of the possible factors inhibiting teamwork and team interaction, in addition to the many other factors such as information overload and uncertainty. However, CSD analysts are often recognized and rewarded based on the attacks he/she has detected and processed. Therefore a notion of "knowledge is power" (and arguably within this case "knowledge is money") is prevalent in this domain, which prevents analysts from

sharing information and knowledge with other analysts leading to minimal collaboration and communication among analysts. Such a disconnection between analysts might have an adverse effect on the performance of them.

Even though we have hypothesized that teamwork is lacking within CSD analyst groups, there is some preliminary evidence from an observational study conducted by Jariwala and colleagues [11] that team work could lead to better performance in cyber defense analysis. However there is a lack of experimental evidence to validate the effectiveness of team work in the cyber defense context. Therefore, in the present lab-based study, we manipulated the effect of teamwork versus group work (control condition) on the performance of cyber defense analysts by priming participants in the teamwork condition with team level rewards that motivate them to work as a team versus priming participants in the group work condition to compete with other analysts for individual rewards.

2 Method

In this study, we are testing the hypothesis that having reward structures which are conducive to team work in CSD analyst groups performing triage level analysis will lead to higher signal detection performance. To test the hypothesis, we conducted a team-based cyber defense analysis experiment. The participants in the experiment used a synthetic task environment [12] called CyberCog [13] to perform the tasks of a cyber defense analyst. Synthetic task environments are simulation environments built to recreate the real world tasks and cognitive aspects of the task with highest fidelity possible, giving less focus towards the appearance of the real world environment [13].

2.1 Simulation Environment

CyberCog is a three-person synthetic task environment that simulates the triage process in cyber defense analysis. The CyberCog system presents a simulated set of network and system security alerts which participants have to categorize as either a benign or suspicious alert based on the analysis they conduct using other simulated information sources such as network and system activity logs, a user database, a security news website, and a vulnerability database. Figure 1 is a screen capture of the CyberCog system where the alerts are presented to the participants. Simulated intrusion alerts used in the system are of 15 different types constructed based on real world intrusion alert types such as alert for malware attack, suspicious email messages, and so forth. However, the alerts used in this system were simplified versions of their real world counterparts to make them understandable for our experimental participants who are not familiar with the domain or the task. Simplified does not imply that the alerts are easy to analyze but simply means that they are presented in a form that is free from technical jargon.

Within the task, participants were trained in depth on 5 of the 15 total alert types. Participants must analyze each alert to decide whether it is a suspicious or a benign

alert by first looking at the corresponding log of the activity that caused the alert. Based on the alert type, the participants also have to leverage more information sources such as employee database or website for further analysis. Some of the alert types are comparatively easier to analyze because they only involve verifying that the user who performed the activity has the authorization to do it or not. Other alerts are more difficult to analyze involving analyzing the source IP address, making judgments on whether sensitive data was transferred, or fusing multiple data to make decisions. However care was taken to make sure that each participant was trained on an equal number of difficult and easy alert types. If a participant was unable to complete an alert due to lack of training or another participant indicated that they were able to solve an alert, the alert could be easily "shared" between the participants.

In summation, the CyberCog system recreates the different aspect of the triage analysis task as it is performed in the real world but in a controlled fashion.

Fig. 1. Screen capture of the web page presenting intrusion alerts in mission 1

2.2 Procedure

Twenty teams comprised of three participants were recruited from the university subject pool to work as CSD analyst teams in the study. 32 were male and 28 were female and gender composition varied across teams. The participants were either given three course credits and a sum of $10 for their participation in the experiment, or four course credits based on participant choice. Participants provided informed consent and were assigned to one of the two conditions: teamwork or group work. The participants were then provided the necessary training for performing the tasks in the experiment. Training was identical between teamwork and group work conditions.

Training. In training, the participants were first given an overview of the cyber domain using a recorded video presentation. They were then provided training to

perform cyber defense analysis tasks using the CyberCog system. They were provided training to be a specialist on analyzing five of the fifteen types of intrusion alerts used in CyberCog system. Each participant received unique individualized training on the five alert types that they were assigned. The training consisted of two sections: A reading section in which the participants would read power point presentation that described how to analyze each type of alert that they were assigned. There was also a hands-on practice section on actually analyzing an example alert of that type.

Experimental Missions. After the training, the participants performed three missions: one practice and two main missions. Practice and the main missions differed by the number of alerts and the time available to complete the mission. In the practice mission, each participant was presented with fifteen alerts and was given fifteen minutes to analyze and classify those alerts. During the two main missions, each participant was presented with seventy-five alerts and was given thirty minutes to analyze and classify those alerts. This meant the participants had to analyze two alerts every minute during the main missions. Thirty minutes was chosen as duration for the main mission to simulate the time crunch and overload that is experienced by an analyst in the real world. The missions were carefully constructed so that each participant would receive an equal number of suspicious and benign alerts to analyze. The missions were also constructed such that each participant would receive a mix of alerts for which the participant received training and alerts for which the participant did not receive any training. During the mission, the participants can either choose to transfer unfamiliar alerts to other members of the team for analysis or learn to analyze those alerts themselves using the lookup system which provided a textual description of the analysis procedure.

During each experiment session, two teams performed the same task in parallel under the same experimental condition. In the team-work condition, the participants were encouraged to work as a team to classify as many alerts as accurately as possible. They were informed that they would be scored as a team (the running scores were presented on a common screen throughout the experimental session) and that all the members of the better of the two teams during that session will receive a reward (i.e., a snack bag). In the group work condition, the participants were instructed to work individually to classify as many alerts, as accurately possible. They were informed that they would be scored individually and that the best of the six participants (two groups of three participants) during that session would receive a reward. The entire experiment session lasted approx. four hours, which included break time between the missions.

Measures. We collected a variety of data and measures from the experiment. Our primary measure of team performance was based on the Signal Detection Theory [14]. For the alerts analyzed the number of hits (number of suspicious alerts the team classified as suspicious), misses (number of suspicious alerts the team classified as benign), false positives (number of benign alerts the team classified as suspicious), and correct rejections (number of benign alerts the team classified as benign) were

recorded. Subjective impressions of workload were measured using the NASA TLX [15] at the end of each mission. A transactive memory measure [16] was administered at the end of the experiment session. For the purpose of this paper we focus on the team performance measure.

3 Results

Hits, misses, false alarms, and correct rejections were collected for each of the 15 alert types completely processed by participants in each team. Although participants received a total of 225 alerts per team, only completed alert classifications were used. That is, during analysis, teams and individuals were not penalized for not classifying an alert. Out of the fifteen alert types, four alerts were identified as harder than the remaining eleven and demonstrated effect of the manipulation. Analysis was completed on these four alert types. All teams were given the chance to process 60 of these "hard" alerts, with the mean processed amount being 42 across conditions. There were no conditional differences on the number of "hard" alerts processed.

The distributions and standard deviations of hits, misses, correct rejections and false positives for "hard alerts" violated the assumptions for d' therefore the non-parametric A' was calculated as a score for signal detection sensitivity. Response bias was calculated using c. A number of SDT scores contained either a 0 or 1 which is not acceptable for A' calculations. A log-linear transformation was conducted on the SDT scores to correct the zeros and ones. A' is the team performance measure in this study. [14]

A mixed-factor 2x2 (condition-between x scenario-within) ANOVA was conducted on A' or sensitivity. A significant effect of condition was found ($F(1,18)$ = 5.662, p = .029). Teams were more sensitive in their classification of alerts (mean = .90) than groups (mean = .85). A' ranges from values 0.5 and 1 with 0.5 indicating lowest performance possible and 1 indicating highest performance possible. Figure 2 shows the comparison of performance between the two experimental conditions.

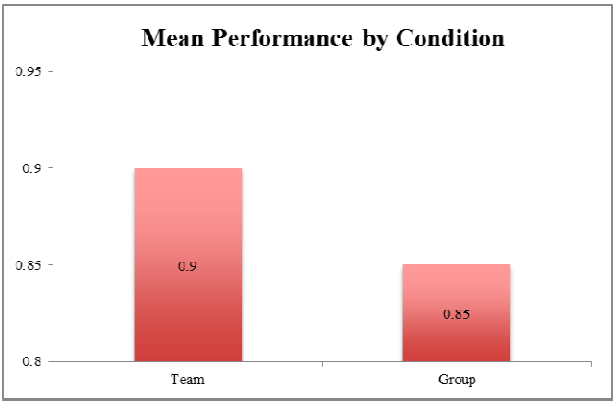


Fig. 2. Mean Performance (A') By Condition

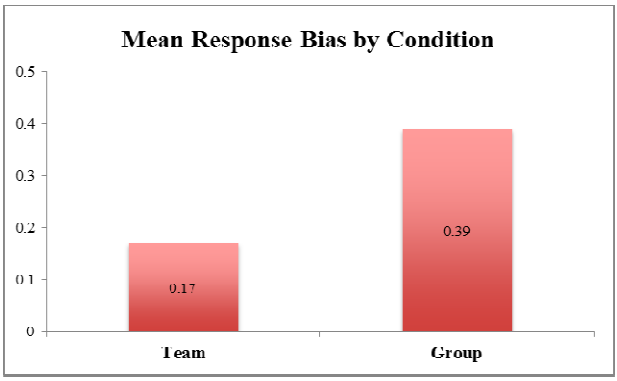


Fig. 3. Response Bias (c) By Condition

In order to examine response bias, we calculated *c*. This measure indicates whether or not the teams were more likely to respond "yes" or "no" when classifying an alert as a suspicious alert or benign. A mixed-factor 2x2 (condition-between x scenario-within) ANOVA was conducted on *c*. Again a significant effect of condition was found ($F(1,18) = 8.756$, $p = .008$). Teams are more conservative (mean = 0.17), that is they are less likely to say an alert is benign when compared to groups (mean = 0.39). Figure 3 shows the comparison of response bias between the two experimental conditions.

4 Conclusion

This study investigated the effect of teamwork versus group work on triage analysis task in cyber defense. SDT measures were collected on alerts processed by teams or groups in order to calculate the team performance measure, A'. As shown in Figure 2, we found that team performance was significantly better than group performance on four "hard" simulated alert types. The participants had to put more cognitive effort into analyzing these four "hard" alert types when compared to other eleven "easy" alert types. The purpose of "easy" alert types was to load the teams or groups with information. To accurately analyze the hard alert types, the specialized training provided prior to the start of mission was necessary. It was difficult for a non-expert to learn and analyze the "hard" alerts during the mission. In contrast, the remaining "easy" alert types were intuitive to non-experts and therefore the groups were able to demonstrate similar performance compared to teams when analyzing "easy" alerts.

Cyber defense analysts in the real world face even more "hard" alerts that are novel, non-intuitive as in zero-day attacks and emerging kinds of threats. They are overloaded by such "hard" alerts than the usual day-to-day kinds of alerts which are often false alarms. It is imperative that analysts analyze such novel, non-intuitive "hard" type of alerts accurately because they are more often the real attack which leads to destructive and expensive consequences.

One might think that putting the extra effort to communicate and collaborate with other team members for analyzing an alert is not essential when one can learn to analyze it by themselves. But as we see from the results that the analysts can achieve higher performance by simply collaborating with other analysts to leverage each other's unique expertise and knowledge to analyze alerts that are novel and non-intuitive to them. Putting the extra effort to analyze all alerts may be detrimental to their performance.

Response biases were calculated to determine if either condition was more inclined to say yes/no in one direction or the other. Both teams and individual groups were more likely to respond with 'no', which given the level of signal to noise is a good indication. We had included noise at five times the level of signal. It becomes interesting to think that teams were less likely to say 'no' than individual groups but still were able to outperform the individual groups. There are any number of speculative reasons: sharing information more readily among the team members, participants were more willing to ask for help since there was no singular incentive, or even the commonality that the team is 'in it together'. Only further investigations could help resolve this presented quandary.

Lastly, the biggest challenge we faced while designing an experiment for studying team cognition in cyber defense has been in building a synthetic task environment with the task and missions that is at the right level of difficulty for the student participants who are not familiar with domain. The task and the missions either get too difficult to understand or so easy that it nullifies external validity of the simulation. However, with the completion of this study we have arrived at a juncture where we have a system that is close to the level we desire. Future work should address this challenge by improving the mission data, task, and also by adding more missions to the STE. Future work should also be investigating more specific cognitive biases that affect team performance in cyber defense analysis.

Acknowledgements. This work has been supported by the Army Research Office under MURI Grant W911NF-09-1-0525. We would like to thank Cliff Wang and our MURI partners for their help and guidance throughout the process of developing CyberCog.

References

1. Ponemon, I.: 2012 cost of cyber crime study. Ponemon Institute, United states (2012)
2. DAmico, A., Whitley, K., Tesone, D., OBrien, B., Roth, E.: Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts. In: Human Factors and Ergonomics Society Annual Meeting Proceedings, pp. 229–233 (2005)
3. D'Amico, A., Whitley, K.: The Real Work of Computer Network Defense Analysts. In: VizSEC 2007, pp. 19–37 (2008)
4. Boyce, M.W., Duma, K.M., Hettinger, L.J., Malone, T.B., Wilson, D.P., Lockett-Reynolds, J.: Human performance in cybersecurity. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 1115–1119 (2011)
5. Liu, P.: Computer-aided Human Centric Cyber Situation Awareness (2009)

6. Champion, M., Rajivan, P., Cooke, N.J., Jariwala, S.: Team-based cyber defense analysis. In: 2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), pp. 218–221 (2012)
7. Salas, E., Dickinson, T.L., Converse, S.A., Tannenbaum, S.I.: Toward an understanding of team performance and training. Teams their Training and Performance, 3–29 (1992)
8. Salas, E., Cooke, N.J., Rosen, M.A.: On teams, teamwork, and team performance: Discoveries and developments. Human Factors: The Journal of the Human Factors and Ergonomics Society 50, 540–547 (2008)
9. Cooke, N.J., Gorman, J.C., Winner, J.L.: Team cognition. In: Handbook of Applied Cognition, pp. 239–268 (2007)
10. Cannon-Bowers, J.A., Salas, E.: Reflections on shared cognition. J. Organ. Behav. 22, 195–202 (2001)
11. Jariwala, S., Champion, M., Rajivan, P., Cooke, N.J.: Influence of team communication and coordination on the performance of teams at the iCTF competition. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 458–462 (2012)
12. Cooke, N.J., Rivera, K., Shope, S.M., Caukwell, S.: A synthetic task environment for team cognition research. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 303–308 (1999)
13. Rajivan, P.: CyberCog A Synthetic Task Environment for Measuring Cyber Situation Awareness (Master's Thesis). Retrieved from ProQuest Dissertations & Theses (PQDT) Database (2011)
14. Stanislaw, H., Todorov, N.: Calculation of signal detection theory measures. Behavior Research Methods 31, 137–149 (1999)
15. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human Mental Workload 1, 139–183 (1988)
16. Lewis, K.: Measuring transactive memory systems in the field: scale development and validation. J. Appl. Psychol. 88, 587 (2003)

Developing Methodology for Experimentation Using a Nuclear Power Plant Simulator

Lauren Reinerman-Jones[1], Svyatoslav Guznov[1],
Joseph Mercado[1], and Amy D'Agostino[2]

[1] University of Central Florida- Institute for Simulation & Training
[2] Nuclear Regulatory Commission (NRC)
lreinerm@ist.ucf.edu

1 Introduction

Many of today's most complicated systems are human-machine systems that involve extensive advanced technology and a team of highly trained operators. As these human-machine systems are so complex, it is important to understand the factors that influence operator performance, operator state (e.g., overloaded, underload, stress) and the types of errors that operators make. Thus, it is desirable to develop an experimental methodology for studying complex systems that involve team operations. This paper looks at Nuclear Power Plant (NPP) operations as a test case for building this methodology. The methodology will reference some aspects/details specific to NPPs, but the general principles are intended to extend to any complex system that involves team operations.

Nuclear Power Plant Operations

NPPs are composed of complex systems that are controlled via a Human System Interface (HSI) located in the Main Control Room (MCR). A minimum of three operators are required to manage and maintain a single nuclear reactor. Two individuals serve as Reactor Operators (RO) and the third is the Senior Reactor Operator (SRO). The types of tasks performed by operators have been classified differently over the years. O'Hara and his colleagues (2008; 2010) spent much time observing the roles of the operators in a NPP and suggest four categories of tasks: Monitoring and Detection, Situational Assessment, Response Planning, and Response Implementation. Monitoring requires checking the plant to determine whether it is functioning properly by verifying parameters indicated on the control panels (Figure 1), observing the readings displayed on screens, and obtaining verbal reports from other personnel. Detection occurs when the operator recognizes that the state of the plant has changed. Situational assessment tasks consist of evaluating current states of NPP systems to determine whether they are within required parameters. Response planning tasks consist of deciding on a plan to diagnose and perform appropriate actions when an event occurs. In NPPs, response planning is largely guided by standardized procedures. The procedures used during accident scenarios, and utilized in the present project, are symptom-based procedures called Emergency Operating Procedures (EOPs). Response implementation tasks consist of performing actions required by response planning (i.e. as directed by the EOP).

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 181–188, 2013.

Response implementation might include selecting a control, performing an action on the control, and watching responses of the system and process resulting from the action.

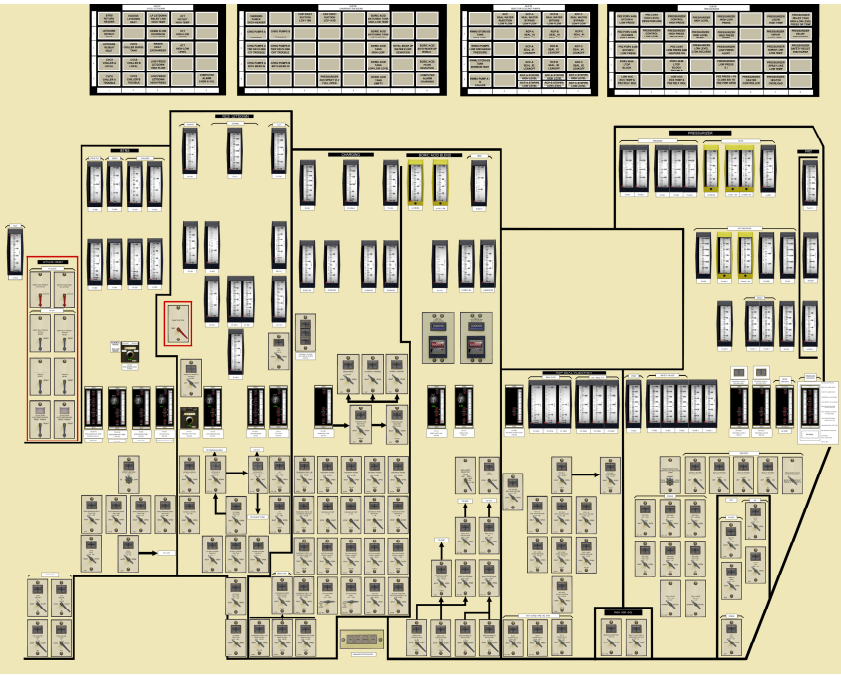


Fig. 1. An example of a NPP MCR control panel

2 Developing a Methodology

Work has been done in the NPP domain to understand the types of tasks operators perform, but systematically investigating and measuring operator performance, errors, and states in a highly controlled experimental setting while executing those tasks has been limited. Developing an appropriate experimental methodology is necessary to effectively evaluate questions concerned with the factors that influence operators' performance, errors, and states.

Test Case

A GSE Westinghouse Four-Loop Pressurized Water Reactor (PWR) simulator will be used in this test case. A non-operator population will serve as participants enabling a larger sample size and reduced cost for experimentation. As the term novice implies, this population has little to no experience with or knowledge of NPPs. Therefore, the environment needs to be simplified, in such a way that will induce participants to experience both the complexity and cognitive requirements incurred by trained operators. In other words, the methodological approach proposed in the present paper

adheres to the principal of different but equal; the populations, EOPs, and control panels are different, but they are different in such a way that is controlled and induces the same level of task demand that would be experienced by each population.

The long-term objective for this work is to examine challenges related to the impact of technology upgrades, automation of tasks, and digital interfaces on the human operators. However, in order to answer those questions, the first step is to begin with exploring the effect that task type has on the workload within each operator role. That is the context within which the below methodology was developed.

Choosing the operating sequence

To reiterate, EOPs are the procedures that operators follow when certain symptoms are present in the plant. These procedures prescribe the type and order of actions that the operating crew takes. For example, if the plant automatically shuts down, operators would enter a procedure called E-0 that would lead them through actions that will diagnose the cause of the shutdown and provide the necessary actions to return the plant to a known safe state. In other domains, where procedures may not be used or not used in the same regimented way, the equivalent may be the event or scenario participants will face (e.g., a hurricane in disaster planning domain). The equivalent is whatever dictates the actions taken by participants. In this case, as the EOP chosen will literally dictate our participants' actions, we are equating EOP with scenario.

Four criteria were established for selecting EOPs best suited for a non-operator sample.

1. Select an EOP that best resembles the typical task flow that operators most commonly face.

A subject matter expert (SME) identified a limited number of frequently used EOPs. A task analysis is being conducted based upon the SME mapping for side-by-side comparison across EOPs. From this mapping, we, along with a SME in NPP operations, will attempt to discern characteristics of a typical task flow. The reason for this criterion is to preserve the fidelity of the task environment by maintaining the typical task flow experienced in a real NPP. Primarily, we want to avoid scenarios that include atypical tasks or order of events as it makes the results less generalizable to other scenarios.

2. Select an EOP that allows the investigation of all roles on the team.

The reason for this criterion was to allow for the assessment of phenomena as relevant to the ROs and the SRO separately, as their primary responsibilities are different. During an EOP, the SRO guides the ROs through symptom-based procedures to identify the events or causes of system alarms, while the ROs interact with control panels to perform actions to alter the state of the NPP. We are interested in understanding the workload associated with different tasks within each role on the team.

3. Select an EOP that requires participants to perform an equal or known ratio of the task types being investigated (e.g., monitoring and detection, and response implementation).

The reason behind this criterion was to enable experimental control of task complexity, which allows for direct comparison between task types. For example, if each operator receives 5 min of monitoring and detection tasking and 5 min of response implementation tasking with each type of tasking followed by administration of the NASA-TLX, then analyses will concretely inform the type of NPP tasks that are more demanding. These results may provide insight into the level of demand to expect when an EOP requires more of one type of task versus another.

4. Select an EOP that incorporates usage of all major categories of instrumentation and controls (e.g., light box, gauge, and switch) within the MCR.

The reason for this is to improve generalizability of experimental results to tasks beyond those carried out in the specific study.

Simplifying the operating environment

In addition to identifying criteria for EOP selection, several steps need to be taken to simplify EOPs and control panels for use with a non-operator sample. The first step is equating the total number of instruments on each panel to provide greater experimental control (Figure 2). This enables us to ensure that performance is not impacted by disparate visual complexity between operators, but, rather the experimental manipulations are the primary factors impacting performance.

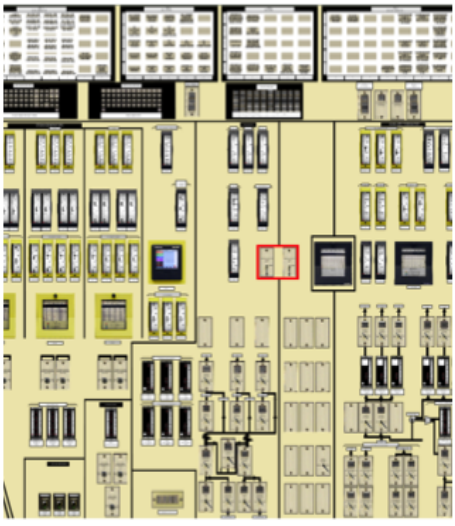


Fig. 2a. Original control panel used by operators

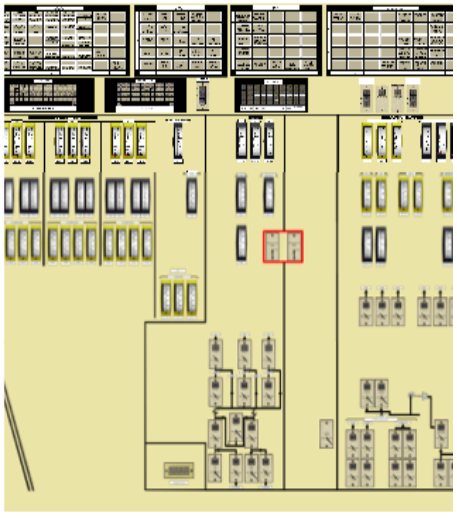


Fig. 2b. Simplified control panel for novice participants

The second step is to modify the EOPs (Figure 3), which refer to gauges and switches by an alphanumeric code, a name, or both, to refer to them only by their alphanumeric code. In addition, all gauges and switches with an alphanumeric code greater than seven characters have been re-coded to a code of seven or less characters,

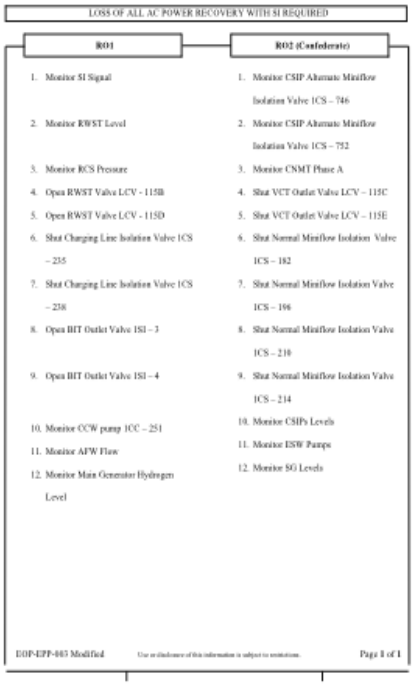
thus yielding a standardized naming convention that maintains the short-term memory principal of seven plus or minus two items. These changes enable simplified communication regarding control board elements. This is important because names of control panel elements are second nature to experts and therefore, are not primary factors that influence operators' performance.

LOSS OF ALL AC POWER RECOVERY WITH SI REQUIRED

Instructions	Response Not Obtained
3. Manually Align SI Valves To Establish SI Injection Mode:	
a. Open CSIP suction from RWST valves: LCV-115B LCV-115D	
b. Shut VCT outlet valves: LCV-115C LCV-115E	
c. Shut charging line isolation valves: 1CS-235 1CS-238	
d. Check RCS pressure - LESS THAN OR EQUAL TO 1800 PSIG	d. WHEN pressure less than 1800 PSIG, THEN do Step 3e. Continue with Step 3f.
e. Check CSIP alternate miniflow isolation valves - SHUT 1CS-746 (Train A CSIP) 1CS-752 (Train B CSIP)	e. Shut the associated block valve: 1CS-745 (Train A CSIP) 1CS-753 (Train B CSIP)
f. Shut normal miniflow isolation valves: 1CS-182 1CS-196 1CS-210 1CS-214	
g. Open SIT outlet valves: 1SI-3 1SI-4	

EOP-EPP-003 | Rev. 20 | Page 6 of 37

Fig. 3a. Portion of original EOP used by operators



LOSS OF ALL AC POWER RECOVERY WITH SI REQUIRED

RO1	RO2 (Confederate)
1. Monitor SI Signal	1. Monitor CSIP Alternate Miniflow Isolation Valve 1CS – 746
2. Monitor RWST Level	2. Monitor CSIP Alternate Miniflow Isolation Valve 1CS – 752
3. Monitor RCS Pressure	3. Monitor CNMT Phase A
4. Open RWST Valve LCV - 115B	4. Shut VCT Outlet Valve LCV – 115C
5. Open RWST Valve LCV - 115D	5. Shut VCT Outlet Valve LCV – 115E
6. Shut Charging Line Isolation Valve 1CS – 235	6. Shut Normal Miniflow Isolation Valve 1CS – 182
7. Shut Charging Line Isolation Valve 1CS – 238	7. Shut Normal Miniflow Isolation Valve 1CS – 196
8. Open BIT Outlet Valve 1SI – 3	8. Shut Normal Miniflow Isolation Valve 1CS – 210
9. Open BIT Outlet Valve 1SI – 4	9. Shut Normal Miniflow Isolation Valve 1CS – 214
10. Monitor CCW pump 1CC – 251	10. Monitor CSIPs Levels
11. Monitor AFW Flow	11. Monitor ESW Pumps
12. Monitor Main Generator Hydrogen Level	12. Monitor SG Levels

EOP-EPP-003 Modified — Use or disclosure of this information is subject to restrictions. — Page 1 of 1

Fig. 3b. Modified EOP used by novice participants

Locating control board elements also requires simplification. When an SRO directs an RO to implement an action using a specific control board element, the SRO will specify the panel element location and in the general region. For example, an SRO might state, "open valve X located on panel A1 in the lower right quadrant." This allows participants to easily locate elements thus reducing the impact that "locating" related issues will have on performance. Once again, as real operators are very familiar with panels, locating elements, generally, does not influence workload or performance.

Finally, real operators complete EOPs that contains tens of steps and will continue operations until the plant returns to a safe and steady-state, the novice participants will complete a fraction of these steps with a defined stopping point. Task type is maintained the amount of steps is reduced, we have kept task types the same. In addition, we wanted to make the duration of work similar to what might be experienced by operators. Although many training sessions can have scenarios that can last up to 3

hours, according to an operations SME, it is not uncommon to see 30-45 minute scenarios especially in initial licensing training. Thus, we thought this a reasonable and realistic starting point for scenario length. Obviously, due to extensive training and frequent practice, experts are able to perform actions more efficiently and effectively and, thus, can do more in less time. We kept have a realism associated with both the type of tasks and duration of work in this study in an attempt to induce similar levels of taskload experienced by operators.

We feel the criteria and simplifications described above, although tailored to the NPP domain, can be used as a starting point for developing experimental methodology for studying complex systems with team operations in other domains.

Selecting Measures

The final stage in the process of developing methodology is selecting measures that allow us to understand performance, determine error types, and understand the state of operators (stressed, overloaded, alert, etc.) while interfacing with complex systems. Performance can be measured in terms of response time, accuracy of actions, and detection of changes. Errors can be categorized along dimensions of slips, lapses, violations, and mistakes. In the NPP context, workload measurement is likely to be important for understanding performance and errors. This assumption is based upon the distinctiveness of the four primary tasks performed by operators. It may be that workload will vary with task type. However, assessing mental workload changes, in this context, may be challenging. No workload measure exists that has been validated in an NPP setting and many subjective assessments interrupt the task or are post-hoc. Interrupting the task changes the overall flow of events and perhaps even the demand requirements of the operators. Questionnaire administration in the middle of a scenario might either hinder operator performance and increase error when the task is resumed or the opposite could occur because a "break" allows the operator to reflect on the scenario event thus far. In comparison, a post-hoc measure might not be sensitive to the dynamic changes occurring in the NPP. The use of physiological metrics assist in circumventing these challenges.

There are many benefits to using physiological metrics as an assessment of mental workload. Most importantly, physiological metrics provide objective and continuous monitoring of the participant's cognitive and physical state (Reinerman-Jones, Cosenzo, & Nicholson, 2010). Several physiological measures are being considered for inclusion in our NPP test case. Electroencephalography (EEG) measures neural activity and is sensitive to changes in mental workload (Figure 4). EEG allows for the continuous monitoring of brain activity without interfering with the primary task (Brookings, Wilson, & Swain, 1996).

Transcranial Doppler (TCD) sonography monitors cerebral blood flow velocity (CBFV) in intracranial arteries and has been commonly used in vigilance studies showing a decrease in CBFV paralleled by decreased performance for sustained attention of highly demanding tasks (Reinerman-Jones, Matthews, Langheim, & Warm, 2010). Vigilance is the detection of infrequent signals amidst non-signals or noise. Much of the operators' responsibility fits the criteria of a vigilance task. Functional Near Infra-Red (fNIR) imaging monitors hemodynamic changes in oxygenated hemoglobin and deoxygenated hemoglobin in the prefrontal cortex (Ayaz et al., 2011).

Fig. 4. An ABM x10 EEG/ECG system worn by a participant

A study by Ayaz et al. (2010) showed that blood oxygenation increases are associated with increasing task difficulty. Electrocardiography (ECG) measures cardiac activity. Heart rate, heart rate variability, and inter-beat interval have been associated with mental workload (Jorna, 1993; Kramer, 1991; Roscoe, 1992, 1993; Veltman & Gaillard, 1996; Wilson, Fullenkamp, & Davis, 1994). Eye tracking measures ocular behavior and can provide insight into task difficulty by providing scan and fixation patterns (Reinerman-Jones, Cosenzo, & Nicholson, 2010).

Awareness of the many possible measures of performance, errors, and states along with understanding the scope and limitations of the operating environment (i.e. simulator capabilities/limitations, physical space, the modified EOPs, required team interaction, and the required actions) enables selecting appropriate assessments.

3 Conclusions

The methodology presented in this paper can serve as a foundation for future human factors testing in the NPP domain and other domains that involve complex systems and team operations. This work will expand understanding of performance in complex systems operations and explain factors, such as new technology or concepts of operation, impact on performance.

Acknowledgment. This work was supported by the Nuclear Regulatory Commission (NRC). The views and conclusions contained in this presentation are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NRC or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

1. Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B.: Optical brain monitoring for operator training and mental workload assessment. NeuroImage 59(1), 36–47 (2011)
2. Ayaz, H., Willems, B., Bunce, S., Shewokis, P.A., Hah, S., Deshmukh, A.R., Onaral, B.: Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. In: Marek, T., Karwowski, W., Rice, V. (eds.) Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations, pp. 21–32. Taylor & Francis Group (2010)
3. Baker, K., Olson, J., Morisseau, D.: Work practices, fatigue, and nuclear powerplant safety performance. Human Factors 36, 244–257 (1994)
4. Brookings, J.B., Wilson, G.F., Swain, C.R.: Psychophysiological responses to changes in workload during simulated air traffic control. Biological Psychology 42(3), 361–377 (1996)
5. Jorna, P.G.A.M.: Heart-rate and workload variations in actual and simulated flight. Ergonomics 36(9), 1043–1054 (1993)
6. Kieras, D.E., Bovair, S.: The role of a mental model in learnng to operate a device. Cognitive Science 8, 255–274 (1984)
7. Kramer, A.F.: Physiological metrics of mental workload: A review of recent progress. In: Damos, D.L. (ed.) Multiple-task Performance. Taylor & Francis Group, London (1991)
8. Mumaw, R.J., Roth, E.M., Vicente, K.J., Burns, C.M.: There is more to monitoring a nuclear power plant than meets the eye. Human Factors 42(1), 36–55 (2000)
9. O'Hara, J.M., Higgins, J.C.: Human-system interfaces to automatic systems: Review guidance and technical bases. Human Factors of advanced reactors (NRC JCN Y-6529) BNL Tech Report No BNL91017-2010 (2010)
10. O'Hara, J., Higgins, J., Brown, W., Fink, R., Persensky, J., Lewis, P., Kramer, J., Szabo, A., Boggi, M.: Human Factors Considerations with Respect to Emerging Technology in Nuclear Power Plants (NUREG/CR-6947). U. S. Nuclear Regulatory Commission, Washington, D.C. (2008a)
11. Reinerman-Jones, L.E., Cosenzo, K., Nicholson, D.: Subjective and objecttive measures of operator state in automated systems. In: Marek, T., Karwowski, W., Rice, V. (eds.) Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations (2010)
12. Reinerman-Jones, L.E., Matthews, G., Langheim, L.K., Warm, J.S.: Selection for vigilance assignments: A review and proposed new directions. Theoretical Issues in Ergonomics Science, 1–23 (2010)
13. Roscoe, A.H.: Assessing pilot workload: Why measure heart-rate, HRV and respiration. Biological Psychology 34(2-3), 259–287 (1992)
14. Roscoe, A.H.: Heart-rate as a psychophysiological measure for in-flight workload assessment. Ergonomics 36(9), 1055–1062 (1993)
15. Veltman, J.A., Gaillard, A.W.K.: Physiological indices of workload in a simulated flight task. Biological Psychology 42, 323–342 (1996)
16. Wilson, G.F., Fullenkamp, P., Davis, I.: Evoked-potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. Aviation Space and Environmental Medicine 65(2), 100–105 (1994)

Modeling Complex Tactical Team Dynamics in Observed Submarine Operations

Tara Smallidge[1], Eric Jones[2], Jerry Lamb[1], Rachel Feyre[2],
Ronald Steed[3], and Abaigeal Caras[1]

[1] Naval Submarine Medical Research Lab, Groton, CT, USA
{tara.smallidge,jerry.lamb,abaigeal.caras}@med.navy.mil
[2] Aptima, Inc., Woburn, MA, USA
{ejones,rfeyre}@aptima.com
[3] UpScope Consulting, Mystic, CT, USA
ronaldsteed@gmail.com

Abstract. Successful submarine operations—those that accomplish the mission while maintaining security and safety—depend on numerous factors. Among the most critical elements driving success are the effectiveness of team behavior and the ability to understand when this behavior breaks down such that this degradation can be mitigated or avoided. While underway, submarine Commanders and other leaders must be attuned and alert to potential precursors that may manifest in decreased performance. This paper describes a framework used to develop performance measures to support formative assessment of team behaviors and to examine team breakdown and degradation. Results are reported from two events: an observation of an operational exercise and a study at the Naval Submarine School concerning the validity and utility of the measures. This preliminary research captured essential aspects of performance and helped define future efforts to develop better tools for assessing team behavior and understanding team breakdown in our warfighters.

Keywords: performance measures, formative assessment, team effectiveness, team breakdown, submarine.

1 Introduction

Submarine crews are operating in an era of emerging complexity in both peacetime and combat operations. Complexity brings with it new, novel, and unpredictable situations which submarine tactical teams must recognize, adapt, and respond quickly and accurately in order to complete the mission or task at hand. The most critical and common complex element is that of understanding tactical team dynamics, performance, and degradation. Maintaining effective operational team performance during prolonged stressful missions is a common challenge faced by the submarine fleet. Naval Submarine Medical Research Laboratory (NSMRL) investigated approaches to further understand the details and characteristics of tactical teams.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 189–198, 2013.

Studies over the past three years at NSRML regarding submarine tactical teams have discerned team behaviors that are aligned with the unique performance needs of submarine warfighters. This research identified and validated five sustainable tactical team practices for submarine crews: Dialogue (interaction among crewmembers), Critical Thinking (how they solve problems), Use of Bench Strength (how they build and utilize all levels of the team), Decision Making (how teams distribute authority to make such decisions), and Problem Solving Capacity (an integration of the other four practices with additional behaviors that measures the degree of tactical complexity that the team can absorb successfully). [1-3]. The five practices are necessary for effective team cohesion and dynamics, and ultimately enable a team to achieve operational resilience. The research method employed a series of workshops (called COMPASS^{SM}, described below) to develop tools that capture behaviors that are aligned with the five practices. The workshops were attended by scientists, engineers, and subject matter experts in the submarine domain. The initial focus of these workshops was to develop, for each practice, a set of Performance Indicators (PIs), i.e., observable and measurable behaviors which allow an instructor or expert to recognize whether a team or individual is performing well or poorly. These PIs were then validated during a three-day observation of a command training exercise performed by an SSGN (cruise missile submarine) crew. The PIs were developed into measurement tools that improve the Submarine Force's ability to assess tactical team behaviors, enhance training through formative feedback, and thus promote successful submarine operations.

Secondly, NSMRL, the Naval Undersea Warfare Center Division Newport, Rhode Island (NUWC) and National Aeronautics and Space Administration (NASA) are proposing to further examine team performance degradation and breakdown of tactical teams during extended missions. While team breakdown is often perceived as a sudden event with a dramatic loss of effectiveness, it may, more appropriately, be viewed as a gradual or incremental process. Therefore, this research is to conduct a set of experiments that measure team performance and determine the relationship (if any) that exists between that performance and a number of variables which may or may not contribute to team performance degradation and, eventually, breakdown. By fully capturing submarine tactical team behaviors and thoroughly understanding the details and specifics of how and when a team breaks down, the Submarine Force will be more capable of resilient action as they encounter increasingly complex combat operations.

2 Measures of Resilient Submarine Tactical Team Behavior

2.1 Overview and Development Process

Prior work performed by NSMRL has identified five team practices that are integral to promoting resilient submarine team operations [1-3]. They focus on interaction among crewmembers (Dialogue), how they work together to solve problems (Critical Thinking), how they build and utilize all levels of the team (Use of Bench Strength),

how the authority to make decisions is distributed among the team (Decision Making), and the degree of tactical complexity that the team can successfully absorb (Problem Solving Capacity). Aptima, Inc.'s COMPASS^{SM} workshop process was employed to identify observable and measurable behaviors that were aligned with these five practices. These behaviors will be used for assessment and as a provision of formative feedback. Initial data collection opportunities validated the initial products of the workshops which included Likert scales, checklists, and narrative descriptions of behavior.

Aptima, Inc.'s COMPASS^{SM} workshop process is a systematic method for identifying essential knowledge and skills and then identifying observable behaviors that provide evidence of that knowledge and those skills at varying levels of expertise. The goal of COMPASS is to develop meaningful and reliable measures that are sensitive to variability in performance and are validated by their relationship to mission outcomes [4]. It does this by combining performance and psychometric theory with extensive subject matter expert input. This input is critical to developing metrics that are firmly tied to the operational domain, are clearly expressed in operationally-relevant terms, and reflect performance at multiple levels within the tactical team (i.e., individual operators, departments, leadership). Leveraging psychometric theory ensures that the resulting measures are reliable, valid, and sensitive to changes in performance (across crewmembers and across time), and that they provide meaningful and diagnostic feedback for post-exercise debriefing. COMPASS has been applied to many complex organizations across the military. In the submarine domain, measures have been developed for routine operations such as coming to periscope depth and weapons employment, while more recent efforts have focused on more interpretive assessments of topics such as Command Team decision making [5].

At the conclusion of the first workshop, approximately 75+ PIs were identified during a thorough discussion of the observable behaviors that an expert observer would expect to see during the course of four representative submarine missions: Intelligence, Surveillance, and Reconnaissance at Periscope Depth, Anti-Submarine Warfare, Strike missions, and Routine Transit. The PIs were continually condensed and refined, and the long list was culled to a much smaller subset of "high-level" PIs that adequately covered the main categories of behavior that were represented. Throughout this process, the PIs were cross-checked with the five practices to ensure that they were aligned with the initial framework. With a reduced set of high-level PIs, the research team then turned to data collection opportunities to begin validating the products of the workshops thus far.

2.2 SSGN Command Training Exercise Observation

In spring of 2012, data were collected during a three-day observation of an SSGN (cruise missile submarine) command training exercise at the Trident Training Facility in Bangor, Washington. The research team divided into two groups of three observers

each, both with a mix of submarine domain and performance measurement experts. Each day was divided into morning and afternoon sessions (eight hours each), with a scheduled two-hour overlap to meet as a group and discuss findings. Each team of three remained on either the morning or afternoon watch for the duration of the multi-day event. During each session, a single person was assigned to a specific practice, and asked to focus on the PIs associated with that Practice. The PI assignments were balanced across the teams and across the days to maximize the amount of data collected for each. All notes were unclassified, and the observers were free to use the previously-developed data-collection sheets in any manner that they considered most useful.

From these data come a number of preliminary findings. The PIs can be developed agnostic of mission, and in fact, they were determined to be nearly agnostic of prior technical knowledge of submarine operations. Even those who were not experienced submariners could pick up on many of the identified behaviors that mark both resilient and brittle teams[1]. The SSGN crew was comprised of three different Watch Sections, i.e., intact teams which are on duty for six hours at a time. In this case, the different Watch Sections provided an excellent opportunity to witness a range of brittle and resilient team behaviors. The abilities of each section seemed to align nicely with this spectrum, and provided first-hand examples of contrasting events that will inform future development of the PIs. The research team determined that there were no "missing" practices, and in fact, some of the five practices (such as Problem Solving Capacity and Decision Making) began to lend themselves to much more richness than previously thought.

With the data that were collected, preliminary analyses were performed that examined the frequency with which each PI was observed. Figure 1 shows the total number of observations of each high-level PI summed across all observers and all days of observation. By far, the most frequently observed high-level PI was "Decision-makers use briefs to build shared understanding." This is not surprising considering that it is an easily observable act that requires someone to communicate verbally with individual or multiple crewmembers. Although all of the high-level PIs focus on observable behaviors, some were more salient than others. For example, discussion and crew engagement can be seen during an exercise, while changes to watch team configuration manifest more slowly and need to be assessed over a longer period of time. Furthermore, some PIs are not frequently observed because they rely on infrequent opportunities to observe them (i.e., the exercise may or may not achieve the necessary conditions for activating a tripwire or pre-planned response.)

[1] Submarine Operational Resilience is a team's capacity to recognize, deep within the command structure, developing danger and opportunity under ambiguous and uncertain conditions. It is a team achievement, requiring conscious and purposeful practices and behaviors. Once a danger or opportunity is recognized, resilient teams are able to adapt and respond in ways that are safe in operations, and bold in war.

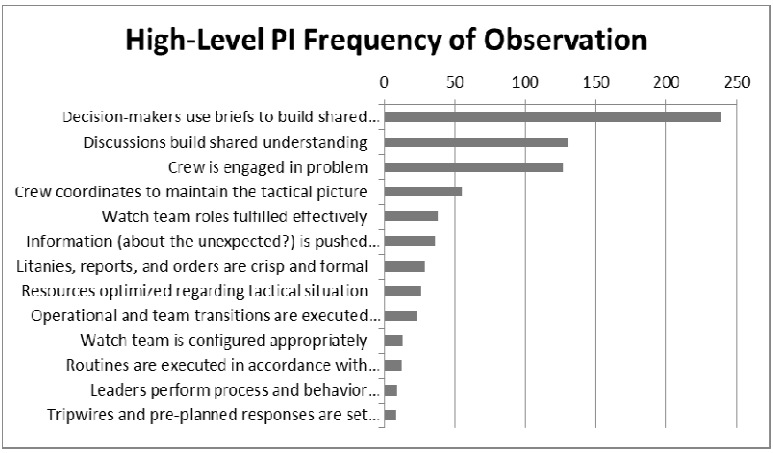


Fig. 1. The total number of observations of each high-level Performance Indicator (PI) during a three-day observation of a SSGN command training exercise

Many of the high-level PIs were associated with multiple practices, if not all of them. Figure 2 below shows the total number of times that each high-level PI was observed through the lens of each practice. Dialogue has the most number of observations, again, because it is the most salient when watching the crew over a short period of time. By contrast, very few of the observed high-level PIs were associated with Problem Solving Capacity or Decision Making. This could suggest that those practices are not yet associated with behaviors readily seen, or are not detectable over the span of a few days of observation. Regardless, data suggest these practices require additional methodological development to capture fully.

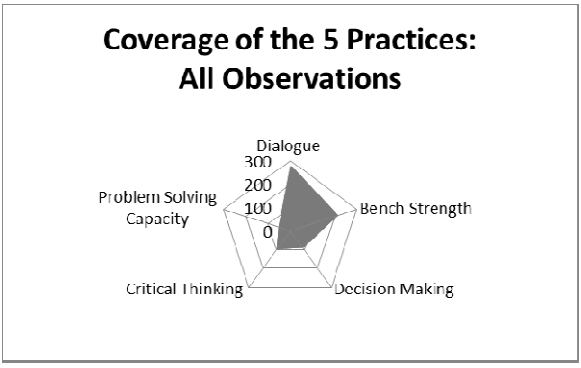


Fig. 2. The total number of times the high-level PIs that are associated with each of the five practices were observed during a SSGN command training exercise (each high-level PI was associated with one or more of the five practices)

2.3 Current State

Following the command training exercise observation, the research team continued to speak with additional experts in the submarine domain to explore ways in which the measures of tactical team behavior could be tailored to meet the needs of its future users. Through these conversations, it became clear that the comprehensive set of measures that were being developed would not be easily integrated into the existing tools and systems that are currently employed by the Fleet to train and evaluate the performance of its warfighters. One key suggestion was to summarize the set of measures so that it could be printed on paper and quickly reviewed by instructors, Commanders, and other leaders to guide their assessments and enable more formative feedback. Therefore the measures were adapted into one-page narrative descriptions of behavior, also referred to as "Team Behavior Maps," that would be much easier to use in this manner.

Each of the five practices (Dialogue, Critical Thinking, Use of Bench Strength, Decision Making, and Problem-Solving Capacity) has its own one-page Team Behavior Map. To develop each of them, the PIs and performance measures for each practice were distilled into a set of observable behaviors that were placed along a continuum of "brittle," "average," and "resilient" behavior. This continuum was divided into five distinct levels of performance that map to this range, and the behaviors were binned into one of the five categories. Observers are then able to assess where a team exists along this continuum of performance by matching observed behaviors to those in each category along the Team Behavior Map scale.

3 Team Performance Degradation and Breakdown

As mentioned earlier, maintaining effective operational team performance during prolonged stressful missions is a common challenge faced by the submarine fleet. While the behavior maps place the teams on a brittle-resilience scale, the ability to assess the change in team performance during increasing stress (complexity) is also necessary. Teams will eventually fail to accomplish their tasking, but such breakdown is preceded by other observable behavior changes. This team degradation and breakdown can be seen during observations of tactical submarine teams when the difficulty of the mission overwhelms their capacity to absorb the complexity. But while team breakdown is often perceived as a sudden event with a dramatic loss of effectiveness, this decrease in performance may in fact be a gradual or incremental process that is presently undetectable. By understanding the specific precursors prior to breakdown, the Submarine Force will be able to design technology and training that more effectively detects and mitigates its impact.

One approach to building an understanding of team degradation and breakdown is to continue gathering data while observing training and at-sea exercises. In addition, NSMRL, in collaboration with NUWCDIVNPT and NASA, propose a series of experiments that will continuously measure team performance while collecting several dependent variables (physiological and behavioral) that are identified as potential early indicators of breakdown. Assessing these variables before and during the

experiment will illustrate when and what variables contribute to team degradation. Another component of this work will research the factors that predict continued effective team performance during prolonged stress. For example, the theoretical underpinnings of this resilience; personality types that are most robust in these environments; the psychological effects of these conditions; crew selection techniques for mitigating breakdown; and, training to build team resilience. Finally, this research aims to gain firm understanding of how a team recovers effective functioning after a breakdown.

A notable factor in studying these types of teams is the flexibility by which they replace losses and augment the team during or after the stressful event; this concept of utilizing reserve capacity is reflected in the practice "Use of Bench Strength," discussed earlier. Although the mission can be accomplished with added support, the original, remaining team must continue functioning and ultimately recover from the breakdown. While stressor types vary between the missions and organizations, varying stress levels should induce team performance degradation or breakdown regardless of the specific mission or organization that is involved. Studying this process requires three critical elements; preliminary data collection, a set of metrics that illuminate precursors to the breakdown, and an operationally realistic environment where the team can perform for days or weeks at a time. Figure 3 below shows two studies that are needed to initially address the team questions that are posed.

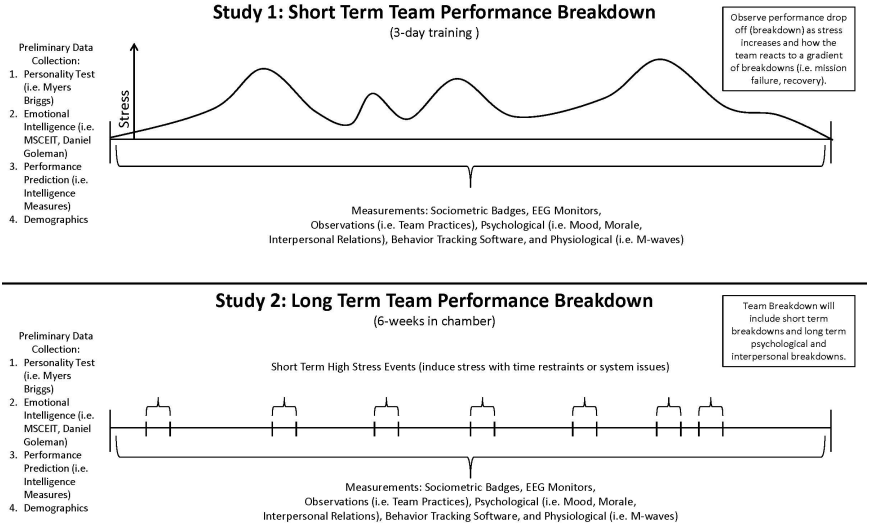


Fig. 3. A sample measure that assesses how a team builds an operational picture and how this assessment may correlate to the proactive transfer of information

In both studies the researchers will collect preliminary data on personality traits (e.g., Myers Briggs); emotional intelligence (e.g., MSCEIT, Daniel Goleman); performance predictors (e.g., Intelligence Measures); and demographics. The researchers plan to correlate these data with team performance degradation and breakdown as

measured through various tools: observer-based measurements (e.g., the measures of Resilient Submarine Tactical Team Behavior), physiological measures (e.g., M-waves), psychological measures (e.g., Mood, Morale, Interpersonal Relations), Electroencephalogram (EEG) monitors, Sociometric Badges (described below), and behavior tracking software. Some of these tools are being used in ongoing research efforts which are examining the behavioral, physiological, and neurological factors that enable a more nuanced assessment of team performance than is currently available.

Several of these ongoing studies provide both insight and techniques for the experimentation that is proposed. For example, the measures of resilient submarine tactical team behavior discussed above potentially provide a sensitive mechanism through which to assess incremental team degradation. Another study with direct relevance is a Defense Advanced Research Projects Agency (DARPA) program with the University of California, Los Angeles (UCLA), Submarine Learning Center (SLC), and NSMRL to wirelessly monitor EEG signals from six submarine team members performing a navigational task. The signals are then time-correlated and assessed using a measure called neurosynchonicity (i.e., a measure of how engaged and collaborative a team is at any given time). This measure has been shown to correlate with scenario events and has been used to identify differences between the ad hoc and mature teams [6].

Physiologically, NSMRL has been conducting at-sea tests of circadian rhythm and lighting by collecting salivary, melatonin, cortisol and alpha-amylase, which are important biomarkers of stress. Specific performance methodologies that have been used include the Multi-Attribute Task Battery, which incorporates tasks that are analogous to activities that aircraft crew members perform in flight, and the Psychomotor Vigilance Task which is a sustained-attention, reaction-timed task that measures the speed at which subjects respond to a visual stimulus.

In addition, NSMRL has the responsibility for testing prospective submariners for suitability for submarine service. The test used, SUBSCREEN (an NSMRL-developed instrument) has been shown to predict losses during the first enlistment. The test is currently being reanalyzed to better predict retention losses. If it is effective, it could possibly become a component of a selection process for effective teams, in addition to standardized test like the Myers-Brigg. Psychological measures will also be included in this study for use in measuring Mood, Morale, and Interpersonal Relations over time. The measures being evaluated during the prolonged stress-induced task will give an indication of how/when a team potentially breakdown in order to provide guidance to our warfighters.

Manually assessing team interactions can, at best, be resource intensive, and for certain team sizes and lengths of time, intractable. Automating such assessments reduces the resources required to do so by decreasing both the number of observers that are required and the time spent manually coding interactions. By removing these constraints, it also increases the amount of data that can be gained because now a group of practically any size can be instrumented to collect data over any length of time. Sociometric Badges (produced by Sociometric Solutions, Inc. [SSI] and the Massachusetts Institute of Technology [MIT]) are small, unobtrusive pieces of hardware that are worn around a person's neck and employ multiple sensors that collect various

types of data as teams of people interact in complex mission environments. The types of data that are recorded include artifacts of speech, face-to-face interactions, body movements, and the proximity of people with respect to one another. In prior experiments the data were analyzed to assess the ability of the Sociometric Badges to automatically and reliably detect behaviors that correlate to team performance [7].

Another tool to assist in behavioral observation and coding is the NASA Behavior Tracking Software developed by Horizon Performance through a Small Business Innovative Research (SBIR) grant. This software, originally designed for the Department of Defense and modified for monitoring astronauts, allows users to track and code human behavior in real time or post hoc using video. The software allows users to timestamp, tag, and rate behaviors as they occur and are observed. These behaviors can then be linked to other data sources. The software can also be used to generate near real-time reports for use by observers or the individuals being observed.

4 Discussion and Next Steps

Team behaviors are crucial to successful submarine operations. If validated, the team practice behavior maps would allow for the accurate evaluation of tactical teams' behaviors, the precious identification of problem areas, and the targeted delivery of formative feedback to communicate, precisely, how a team can improve its resilience. Validation will require disparate observers who are using the tools to record similar assessments (reliability), descriptions that sufficiently capture the range of behaviors that define the team's performance (sensitivity), as well as accurate correlations between Team Behavior Map rankings and team performance (validity). If the Behavioral Practices are diagnostic of resilient submarine team behavior, then higher scores should correlate with successful performance. The Team Behavior Maps will be further validated by comparing the assessments of the observers who are using them (i.e., members of the research team, Navy personnel and/or contractors). If the worksheets are a reliable assessment tool, the individuals' ratings should be consistent. The Team Behavior Maps will be instrumental in future research efforts, such as the proposed experimentation to examine team breakdown and degradation.

Team breakdown and degradation is often perceived as a sudden event with a dramatic loss of effectiveness, however the breakdown of a team may in fact be a gradual, observable process. As is typically seen in the submarine domain, effective operational team performance is difficult to maintain, especially during prolonged stressful missions. By understanding the proposed theoretical underpinnings of the effect these missions have on teams—including the personality types that are most robust in these environments, the psychological effects of these conditions, crew selection techniques for mitigating breakdown, and training to build team resilience—the Submarine Force will be able to design technology and training that more effectively detects and mitigates their impact.

Overall, this program of research will assist the Submarine Force as they encounter the increasing complexity of combat operations, serve to improve the individual and team selection and screening process, and evaluate intact teams for potential vulnerabilities such that they can be trained to be more resilient when faced with these challenges.

Acknowledgements. This work was completed under a contract with the Naval Submarine Medical Research Laboratory (NSMRL). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSMRL. We gratefully acknowledge the support and assistance of the crew that participated in the study, the Trident Training Facility (Bangor, WA), the Naval Submarine School (Groton, CT), and the Naval Undersea Warfare Center Division Newport.

References

1. Steed, R., Lamb, C., Severinghaus, R.: Submarine Operational Resilience: Team Performance Behaviors during Underway Tactical Operations. Paper presented at the National Defense Industrial Organization Joint Undersea Warfare Technology Conference, Groton, CT (September 2011)
2. Lamb, C., Steed, R., Severinghaus, R.: Improving Mission Effectiveness by Increasing Team Problem-Solving Capacity. Paper presented at the National Defense Industrial Organization Joint Undersea Warfare Technology Conference, Groton, CT (September 2012)
3. Steed, R., Lamb, C., Severinghaus, R., Lamb, J., Fields, E., Caras, A.: Team Performance Behaviors during Underway Tactical Operations (NSMRL//TR–2012-0001). Naval Submarine Medical Research Laboratory, Groton (2012)
4. MacMillan, J., Entin, E.B., Morley, R.M., Bennett Jr., W.R.J.: Measuring team performance and complex and dynamic military environments: The SPOTLITE method. Military Psychology (in press)
5. Jones, E., Steed, R., Diedrich, F., Armbruster, R., Jackson, C.: Performance-based metrics for evaluating submarine command team decision-making. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) FAC 2011, HCII 2011. LNCS, vol. 6780, pp. 308–317. Springer, Heidelberg (2011)
6. Stevens, R., Galloway, T., Wang, P., Berka, C., Tan, V., Wohlgemuth, T., Lamb, J., Buckles, R.: Modeling the NeurodynamicComplexity of Submarine Navigation Teams. Computational & Mathematical Organization Theory 18, 1–24 (2012)
7. Jones, E., Lansey, J., Kern, D., Whetzel, J., Diedrich, F., Haass, M.: Automated assessment of submarine team performance using Sociometric Badge technology. Paper presented at the National Defense Industrial Organization Joint Undersea Warfare Technology Conference, Groton, CT (September 2012)

How Tasks Help Shape the Neurodynamic Rhythms and Organizations of Teams

Ronald Stevens[1], Trysha Galloway[1], Gwendolyn Campbell[2], Chris Berka[3], and Pierre Balthazard[4]

[1] IMMEX/UCLA, The Learning Chameleon, Inc., Los Angeles, CA, USA
[2] NAVAIR, Orlando, FLA, USA
[3] Advanced Brain Monitoring, Inc., Carlsbad, CA, USA
[4] St. Bonaventure University, School of Business, Allegany, NY, USA
immex_ron@hotmail.com
trysha@teamneurodynamics.com
campbellge@navair.navy.mil
chris@b-alert.com
pbalthaz@sbu.edu

Abstract. We have modeled neurophysiologic indicators of Engagement and Workload to determine the influence the task has on the resulting neurodynamic rhythms and organizations of teams. The tasks included submarine piloting and navigation and anti-submarine warfare military simulations, map navigation tasks for high school students and business case discussions for entrepreneurial / corporate teams. The team composition varied from two to six persons and all teams had teamwork experience with the tasks. For each task condition teams developed task-specific neurodynamic rhythms. These task-specific rhythms were present during much of the task but could be interrupted by exogenous or endogenous disturbances to the team or environment. The effects of these disturbances could be rapidly detected by changes in the entropy levels of the team neurodynamics symbol streams. These results suggest the possibility of performing task-specific comparisons of the rhythms and organizations across teams expanding the opportunities for rapid detection of less than successful performances and targeted interventions.

Keywords: team neurodynamics, entropy, coordination dynamics, rhythms.

1 Introduction

Teamwork is an important, and most would argue, integral part of all human activities. Like most forms of social coordination, teamwork is not simple. Early studies showed that communication is dynamic during social interactions like teamwork with cyclic exchanges having both synchronous and lead-lag relationships; when repeated these can evolve into shared rhythms and refined speech patterns [1]. It is now widely appreciated that within the context of coordinated team activity such linkages and synchronizations extend beyond speech to include gestural, postural,

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 199–208, 2013.

functional, and physiologic systems [2-4]. It is not surprising that neurophysiologic events are the underpinnings of these dynamics yet it is only recently that their evolving dynamics in real-world teamwork settings have begun to be modeled [5-9].

Our work has focused on developing an information and organization-centric framework for team neurodynamics that is information centric in the sense that raw EEG measures from each team member are combined into symbols showing the levels of different cognitive measures of each team member and the team as a whole [10, 11]. These neurodynamics symbol streams (NS) are probed for regions containing information related to team performance much in the way that words in a sentence or the codons in nucleic acids convey information. Importantly, fluctuations in the mix of symbols identify 'interesting periods' of team organization and the frequency, duration, and magnitude of these fluctuations can be quantified by measuring the Shannon entropy of the data stream [12].

The purpose of this study was to expand a research framework describing successful teamwork by focusing on how elements of the task help shape team neurodynamics. This perspective could be useful for better understanding team-related concepts like organization, rhythm, resilience and the effects of exogeneous and endogeneous disturbances to the team, and lead to the development of more quantitative approaches for comparing across teams. To test the generality of this approach we describe the team neurodynamics of four tasks where the teams were experienced with the task and had worked with the other members of the team, i.e. in the Phase 4 of Team Development as described by Kozlowski et al [13].

2 Hypotheses

The hypotheses for this study were:

1. Teams develop identifiable task-related neurodynamic rhythms and organizations
2. These rhythms and organizations are dynamically modified in response to endogenous and exogenous disturbances to the task

3 Methods

3.1 Tasks and Participants

Map Navigation Task (N = 15 High School Teams)

The task was a two-person problem solving / navigation exercise based on the Edinburgh Map Task corpus [14]. Two team members sat facing and each had a sketch-map with several landmarks on it. The two maps were similar, but not identical and they could not see each other's map. One person, the instruction giver (Giver or G), had a path printed on the map and attempted to verbally guide the other person, the instruction follower (Follower or F) in drawing that path on the Follower's map. The subjects for this task were fifteen 11^{th} and 12^{th} grade science student teams.

Anti-Submarine Warfare Helicopter Teams (N = 3)
The second task was a training exercise for experienced Anti-Submarine Warfare Helicopter Teams (ASWT). Three crewmembers, the pilot, the sonar operator and the tactical officer performed simulated search, track and attack missions in support of surface combat groups. The role of the pilot was to steer the helicopter to the location of a submarine sighting and to fly appropriate paths for buoy configuration. When the approximate location was reached the sonar operator directed the three-dimensional positioning of the passive and active sonar buoys. The tactical officer directed the overall mission and munitions drop. There were three teams based out of Orlando, FLA and San Diego, CA that each conducted two mission simulations; these teams had in-flight crew experience.

Submarine Piloting and Navigation Teams (N = 21)
Submarine Piloting and Navigation (SPAN) is a high fidelity simulation where events include encounters with approaching ship traffic, the need to avoid shoals, changing weather conditions, and instrument failure [15]. Each SPAN session contains three segments beginning with a Briefing where the overall goals of the mission are presented. Next, the Scenario is a dynamically evolving task containing both easily identified and less well-defined processes of teamwork. The final segment, the Debrief is the most structured part with team members reporting on their performance.

Entrepreneurial Teams (N = 6)
A fourth set of data was collected from teams of experienced / advanced student entrepreneurial teams at two international business schools. The simulations lasted ~40 minutes and were structured around business case discussions of corporate social responsibility concerns [16]. The task segments included: 1) defining the task and surfacing pertinent information; 2) prioritizing and discussing issues; 3) developing a team consensus about how to proceed; and 4) formalizing the team recommendation.

3.2 Electroencephalography (EEG)

The B-Alert® system by Advanced Brain Monitoring, Inc. contains an easily-applied wireless EEG system that includes software that identifies and eliminates multiple sources of biological and environmental contamination and allows second – by - second classification of cognitive state changes [17]. The 9-channel wireless headset includes sensor site locations: F3, F4, C3, C4, P3, P4, Fz, Cz, POz in a monopolar configuration referenced to linked mastoids. B-Alert® software acquires the data and quantifies engagement (EEG-E) and mental workload (EEG-WL) in real-time using linear and quadratic discriminant function analyses with model-selected PSD variables in each of the 1-hz bins from 1 – 40 Hz, ratios of power bins.

3.3 Team Neurodynamics

When combined data from multiple time series (i.e. team members) are treated as symbols instead of numeric points it becomes easier to mine them and detect interesting patterns. Normalized second-by-second values of EEG-E or EEG-WL were concatenated into vectors representing the levels being expressed by each team member. For instance, in Fig.1A team members 1, 3 and 5 were expressing below average levels of EEG-E and were assigned values of -1. Team members 2 and 4 were expressing average levels of EEG-E and were assigned the value 1, and team member six was expressing above level values and was assigned the value 3; the vector representation was therefore (-1, 1,-1,1,-1,3). Using artificial neural networks (ANN), the vectors from multiple performances were modeled into collective team variables termed neurodynamic symbols of engagement (NS_E) or workload (NS_WL). ANN classification of these second-by-second vectors created a symbolic state space showing the possible combinations of either EEG-E or EEG-WL across team members (Fig. 1B). Experimentally, the EEG data has been modeled into state spaces between 9 and 900 symbols depending on the task and team [17].

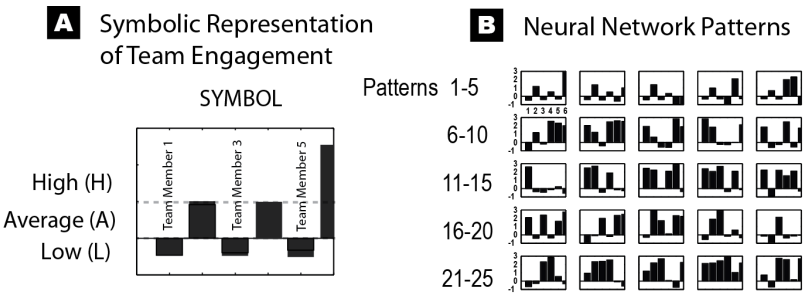


Fig. 1. Data Flow for Creating Team Neurodynamics Models. ANN classification of second-by-second vectors (A) creates a symbolic state space showing the possible combinations of EEG-E or EEG-WL across members of the team (B).

While a symbolic representation of the team state is useful for characterizing team neurodynamics, it is not the best representation for quantifying team neurodynamics. Although there are methods for the quantitative representation of symbols, we chose a moving average window approach to derive numeric estimates of the Shannon entropy of the NS symbol stream. Entropy is expressed in terms of bits; the maximum entropy for 25 randomly-distributed NS symbols would be log2 (25) or 4.64. For comparison, an entropy value of 3.60 would result if roughly half (12) of the NS symbols were randomly expressed. To develop an entropy profile over a session, the NS Shannon entropy was calculated at each epoch using a sliding window of the values from the prior 60 -100 seconds. As teams entered and exited periods of organization, the entropy should fluctuate as a function of the number of NS symbols being expressed by the team during a block of time [15]. Entropy is a quantity, the value of which is determined by the state of the system, in our case with regard to the EEG-E or EEG-WL of the team members. By itself, it says nothing about the state of the system; this information comes from the NS symbols.

4 Results

4.1 Map Task

The detailed NS_WL dynamics for one Map Task team are shown in Fig. 2. NS symbol 2 was expressed twice as often as other symbols (Fig. 2A), represented periods where the Giver expressed high levels of EEG-WL and the Follower was expressing average or, below average levels. The dominance of NS 2 was also seen in the second-by-second NS symbol expressions (Fig. 2B). Around epoch 200 the Follower began having difficulties drawing the map with the mouse. As the difficulties persisted (indicated by the frequency of mouse clicks in Fig. 2D) this resulted in a team reorganization where NS 2 expression was sequentially replaced by NS 4 (G↕F↑), NS 5 (G↓F↑), NS 7 (G↓F↕) and NS 9 (G↓F↓); i.e. the team slowly reduced its EEG-WL. This increased organization was reflected in the slowly decreasing entropy levels. Once the Follower regained control of the mouse the entropy levels rapidly increased as the team re-established its normal operating rhythm NS 2.

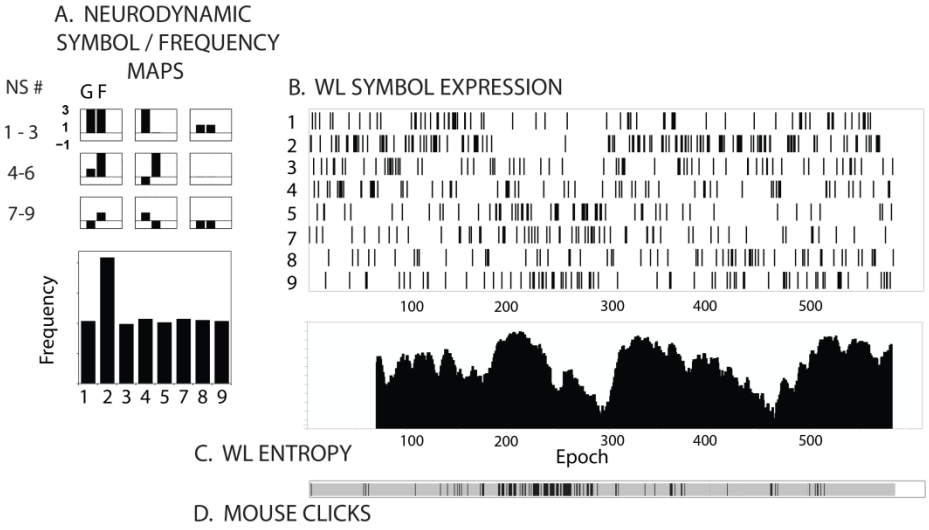


Fig. 2. Linking NS symbols (A) with temporal NS expressions (B), entropy fluctuations (C) and drawing mouse clicks (D)

4.2 ASWT Teams

The neurodynamics are shown for one ASWT team where the major task segments Search, Track and Attack have been identified. For these teams, across team NS-WL models were developed by pooling the NS vectors from four performances and then testing teams individually against this model [15]. The NS maps for EEG-WL showed that that NS 1 and 25 had twice the expression of the remaining symbols. These symbols represent periods where the ATO & SO had high EEG-WL levels and

the Pilot had low (i.e. [ATO,SO]↑P↓) (e.g. NS 1) or the combination [ATO,SO] ↓P↑ (e.g. NS 25). From the perspective of teamwork these NS_WL patterns are consistent with what would be expected from the task as the ATO & SO work closely together once contact is made while the pilot needs less second-by-second coordination with the other members while flying to the initial location, or when changing the search area. Entropy fluctuations were present in the three major task segments that corresponded to identifiable simulation events like in Fig. 3C.1.a where the sonar instrument was malfunctioning and needed repair. During that period the predominant NS_WL symbols were NS 3-10 indicating periods where all team members had average or below average EEG-WL.

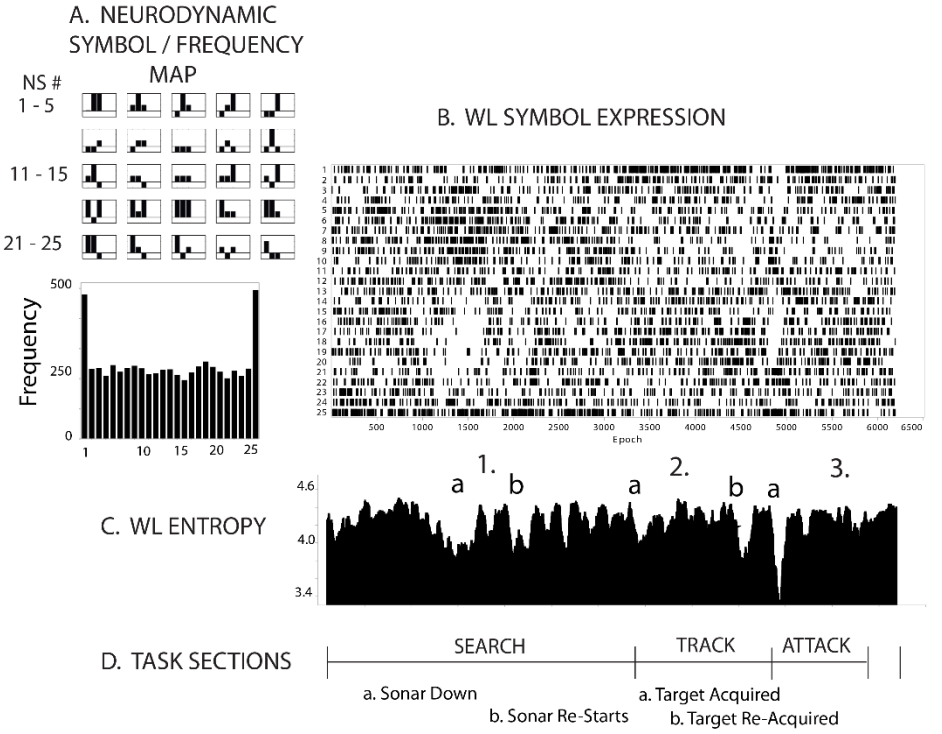


Fig. 3. Linking NS symbols (A) with temporal NS expressions (B), entropy fluctuations (C) and segments of the task (D)

4.3 Submarine Piloting and Navigation Teams

The members of SPAN teams also have defined roles but with up to ten team members the teamwork is more complex. As with other teams, the NS expressions were not uniform, but showed qualitative changes over time, particularly at the Scenario / Debriefing junction. For instance NS_WL symbols 10, 11 and 18 which were poorly expressed during the Scenario, dominated during the first half of the Debriefing. Qualitative dynamic changes also occurred during the Scenario, but these

were generally less obvious than those at task junctions. They were sufficient however to be detected by entropy fluctuations such as those between epochs 2300 - 2500 when the submarine deviated from its safe operating envelope (Fig. 4C.2.c). More pronounced neurodynamic reorganizations were seen during the Debriefing Segment (Fig. 4C.3.a) as the causes for this deviation were discussed.

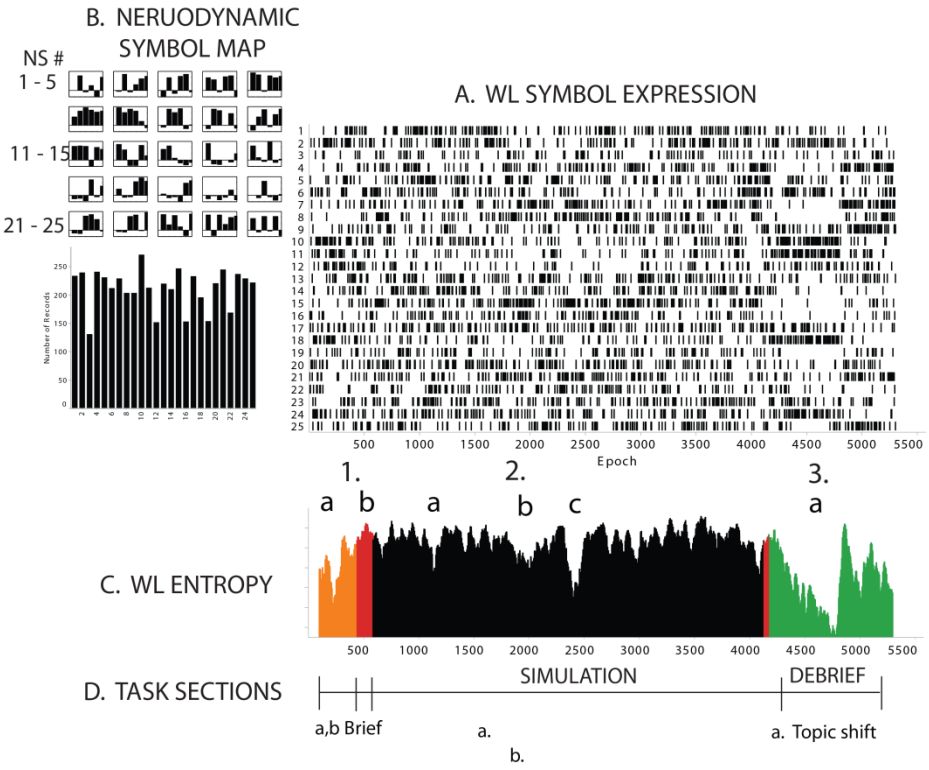


Fig. 4. Linking NS symbols (A) with temporal NS expressions (B), entropy fluctuations (C) and segments of the task (D)

4.4 Entrepreneurial Teams

Teams participating in the business simulations do not have defined member roles like other tasks and the discussions and teamwork are less structured. The data of one team is shown in Fig. 5 where the second-by-second expression of the 25 NS (Fig. 5A) are plotted (Fig. 5B) along with the profile of the Entropy (Fig. 5C). Sections of the entropy profile have been highlighted to indicate task segments. To show the modeling generality this study highlights EEG-E rather than EEG-WL.

Prior to the start (Fig 5C.1) many of the team members had low EEG-E (NS 6, 13, & 14) as general instructions were given. The team then began to surface issues and during this segment NS 1 emerged as the dominant symbol. This symbol represented periods where team member 1 had high EEG-E while the rest were average / low.

This team rhythm intensified as the end of this session was approached and the resulting organization was reflected in a drop in the entropy. The team was then instructed to begin developing a consensus and NS 1 was replaced by a variety of other NS. After ~10 minutes NS 1 re-emerged as the dominant NS as consensus was reached. The team then entered the last segment of the task where their recommendations were finalized (Fig. 5C.4).

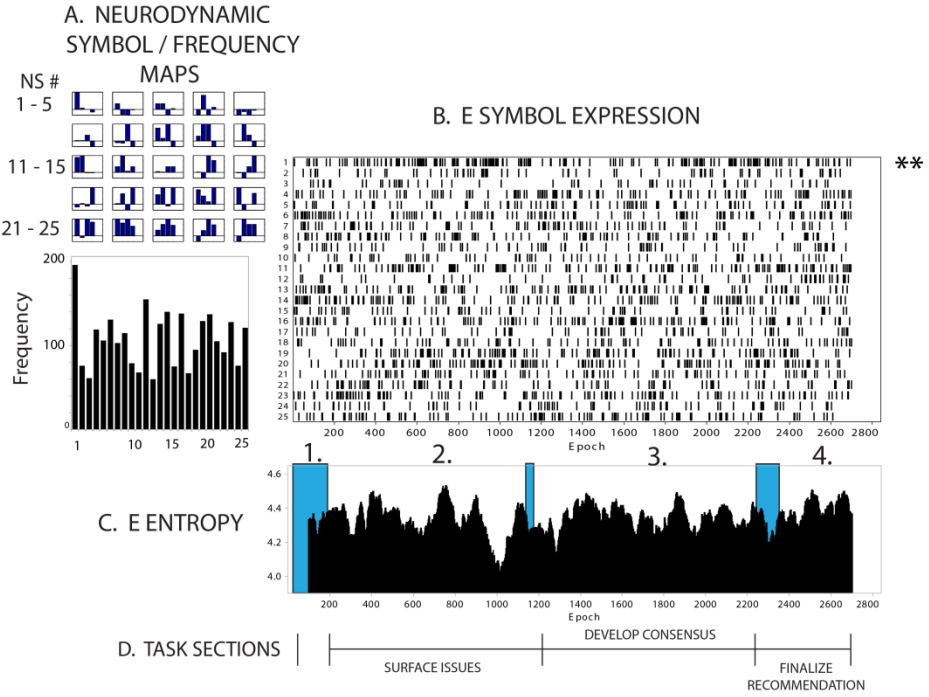


Fig. 5. Linking NS symbols (A) with temporal NS expressions (B), entropy fluctuations (C) and segments of the task (D). The blue regions indicate times where instructions were being given. The ** represents the dominant symbol.

5 Discussion

The first hypothesis was that teams develop task-related neurodynamic rhythms and organizations when performing a task. As cycles and rhythms are widespread across different systems and subsystems during social interaction it would not be unusual to find a form of neurodynamics rhythm. What was less certain was if and how these rhythms would be manifested in EEG-defined measures of Engagement or Workload, and what the prevalence, magnitude, and duration of such rhythms would be. The data in this study suggest that many, if not most successful teams develop what we would term a Normal Operating Rhythm (NOR). The NOR is operationally described as a symbolic representation of a quantitative combination of an EEG-defined measure

that is expressed most often. In terms of complexity theory these preferred patterns of neurodynamics expression can be thought of as a rhythm that the team frequently returns to or an attractor system. As described by Goldstein et al [18] attractors are likely more than repetitive patterns, but are more representative of the underlying system of beliefs '…the core drivers of organizational culture that lead to consistent individual choices and actions.'

These rhythms often did not appear immediately, particularly with teams that had not worked together, but emerged with the progress of the performance. For both the MT and the ASWT teams these rhythms and organizations were not team or session specific but were seen across teams and sessions indicating a more generalized organizational phenomena. Such rhythms may be useful for evaluating different combinations of team members to determine which teams develop the most efficient and effective synchronies.

Neurodynamic re-organizations were often a result of these rhythms or of disturbances to the rhythms, but the question remains open as to why such organizations develop. A simple answer would be that it is an energy savings / efficiency device, i.e. self-organization of complex systems often results in reduced system entropy. When one complex system (task) interacts with a second complex system (team) it is difficult to reduce the constraints of the task, but the degrees of freedom of interaction of the team members can be reduced by mutually agreeing on a defined protocol of exchanging information. A final possibility is that they are a manifestation of shared situation awareness or of team macrocognition. If so, they may provide a pathway for linking the neurodynamic and behavioral models of teamwork.

Acknowledgements. This work was supported in part by The Defense Advanced Research Projects Agency under contract number(s) W31P4Q12C0166, and NSF SBIR grants IIP 0822020 and IIP 1215327. The views, opinions, and/or findings contained are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

References

1. Gottman, J.M.: Detecting cyclicity in social interaction. Psychological Bulletin 86, 338–348 (1979)
2. Ashenfelter, K.: Simultaneous analysis of verbal and nonverbal data during conversation: symmetry and turn-taking. Unpublished thesis. University of Notre Dame (2007)
3. Shockley, K., Santana, M.-V., Fowler, C.A.: Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance 29(2), 326–332 (2003)
4. Gorman, J.C., Amazeen, P.G., Cooke, N.J.: Team coordination dynamics. Nonlinear Dynamics, Psychology, and Life Sciences 14, 265–289 (2010)

5. Stevens, R.H., Galloway, T., Berka, C., Sprang, M.: Can neurophysiologic synchronies provide a platform for adapting team performance? In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) FAC 2009. LNCS, vol. 5638, pp. 658–667. Springer, Heidelberg (2009)
6. Stephens, G., Silbert, L., Hasson, U.: Speaker-listener neural coupling underlies successful communication. Proc. Nat. Acad. Sci., http://www.pnas.org/cgi/doi/10.1073/pnas.1008662107
7. Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., Garnero, L.: Inter-brain synchronization during social interaction. PlosOne 5(8), e12166 (2010), doi:10 1371/journal.pone0012166
8. Dodel, S., Cohn, J., Mersmann, J., Luu, P., Forsythe, C., Jirsa, V.: Brain signatures of team performance. In: Proceedings HCI International 2011, Orlando, FLA (2011)
9. Berka, C., Levendowski, D.J., Cvetinovic, M.M., Petrovic, M.M., Davis, G., et al.: Real-Time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. International Journal of Human-Computer Interaction 17(2), 151–170 (2004)
10. Stevens, R.H., Gorman, J.C., Amazeen, P., Likens, A., Galloway, T.: The organizational dynamics of teams. Nonlinear Dynamics, Psychology and Life Sciences 17(1), 67–86 (2013)
11. Stevens, R.H., Galloway, T., Wang, P., Berka, C.: Cognitive neurophysiologic synchronies: What can they contribute to the study of teamwork? Human Factors 54, 489–502 (2012)
12. Shannon, C., Weaver, W.: The mathematical theory of communication. University of Illinois Press, Urbana (1949)
13. Kozlowski, S.W.J., Watola, D.J., Nowakowski, J.M., Kim, B.H., Botero, I.C.: Developing adaptive teams: A theory of dynamic team leadership. In: Salas, E., Goodwin, G.F., Burke, C.S. (eds.) Team Effectiveness in Complex Organizations: Cross-disciplinary Perspective and Approaches. SIOP Frontier Series. LEA, Mahwah (2009)
14. Doherty-Sneddon, G., Anderson, A., O'Malley, C., Langton, S., Garrod, S., Bruce, V.: Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. Journal of Experimental Psychology: Applied 3(2), 105–125 (1997), doi:10.1037/1076-898X.3.2.105
15. Stevens, R., Galloway, T., Wang, P., Berka, C., Tan, V., Wohlgemuth, T., Lamb, J., Buckles, R.: Modeling the Neurodynamic Complexity of Submarine Navigation Teams. Comput. Math. Organ Theory (2012), doi:10.1007/s10588-012-9135-9
16. Pless, N., Maak, T.: Levi Strauss & Co: Addressing child labour in Bangladesh. In: Mendenhall, M.E., Oddou, G.R., Stahl, G.K. (eds.) Readings and Cases in International Human Resource Management and Organizational Behavior, 5th edn., pp. 446–459. Routledge, London (2012)
17. Stevens, R.H.: Charting neurodynamics eddies in the temporal flows of teamwork. In: Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting (October 2012)
18. Goldstein, J., Hazy, J., Lichtenstein: Complexity and the nexus of leadership. Palgrave, Macmillan (2010)

Neurophysiological Estimation of Team Psychological Metrics

Maja Stikic[1], Chris Berka[1], David Waldman[2], Pierre Balthazard[3], Nicola Pless[4], and Thomas Maak[4]

[1] Advanced Brain Monitoring, Inc., Carlsbad, CA, USA
[2] Arizona State University, W.P. Carey School of Business, USA
[3] St. Bonaventure University, School of Business, USA
[4] ESADE Business School, Ramon Llull University, Spain
{maja,chris}@b-alert.com, waldman@asu.edu, pbalthazard@sbu.edu, {nicola.pless,thomas.maak}@esade.edu

Abstract. The goal of this study was to explore the feasibility of continuous neurophysiological assessment of different psychological aspects of a team process. The teams consisted of the MBA students who discussed and attempted to solve a case problem dealing with corporate social responsibility (i.e. child labor). At the end of the team process, two types of psychological metrics (i.e., engagement and leadership) were assessed by team members, both at the individual and team levels. These metrics showed significant correlations with the team performance scores derived by four trained coders. Two of them rated the teams' solutions in terms of effective problem solving, decisiveness, and creativity. The other two coders rated the level of moral reasoning displayed in the solutions. The psychological metrics were then estimated based on quantitative electroencephalography (qEEG). Different modeling techniques, such as linear and quadratic discriminant function analysis (DFA) and linear regression were applied to the processed qEEG data. The models were evaluated through auto-validation, but also through cross-validation to test stability of the models in the team-independent training setting. The experimental results suggested that qEEG could be effectively used in the team settings as an estimator of individual and team engagement, as well as the leadership qualities shown by team members. Our findings suggest that qEEG can help in understanding, and perhaps building, optimal teams and team processes.

Keywords: team process, engagement, leadership, electroencephalography.

1 Introduction

There is a growing interest in studying different aspects of collaborative teamwork such as engagement [1] and leadership [2] due to evolving task demands and the necessity of building optimal teams that accomplish the tasks successfully and effectively. Psychological metrics are typically measured using traditional psychometric assessment methodologies at the conclusion of a team process. In [3] it has been

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 209–218, 2013.

argued that neuroscience can provide more ecologically-valid assessment of psychological metrics. Recent advances in the technical design of the qEEG hardware and software platforms enable practical application of qEEG in studying team processes. The main advantage of the qEEG-based team assessment is that it is continuous, and it does not require disruption of the ongoing team process. For example, [4] utilized the qEEG data for modeling team dynamics in complex military tasks.

This paper explores qEEG-based estimation of engagement and leadership at the individual and team levels during the team discussion of a social responsibility case problem. The proposed approach is not computationally expensive, it accounts for individual variability inherent in the qEEG data, and it leverages a general trend of qEEG changes to characterize individual and team engagement and leadership. Unlike [5], where transformational leaders were classified based on the qEEG data during the resting eyes closed session, we focus on the more challenging team setting.

2 Materials and Methods

In this section, our study protocol is outlined, the assessed performance and psychological metrics are introduced, and the qEEG acquisition system is described together with signal processing and data analysis.

2.1 Study Protocol

The students at a business school in Europe formed 31 teams of either 4 or 5 individuals. The overall sample comprised 146 students with the mean age of 28.7 years. The participants were ethnically diverse (61.5% were Caucasian, 20.7% were Asian, and 15.6% were Hispanic) and gender balanced (64.4% were males).

Each subject first completed a sustained attention task (3-choice active vigilance task - 3CVT) that required subjects to discriminate one primary target (presented 70% of the time) from two secondary non-target geometric shapes that were randomly interspersed over a 20 min period. Participants were instructed to respond as quickly as possible to each stimulus. A brief training period was provided prior to the start of the task to minimize the practice effects.

The problem solving task addressed a corporate social responsibility case of the Levi Strauss Company involving child labor issues in Bangladesh [6]. Each team member was first given approximately 40 min to read the case individually, consider the issues presented in the case, and form initial solutions, which were typed into computer files by respective students. Afterwards, the subjects were engaged in the team discussion process with the goal to derive a common solution to the case. The discussion lasted up to 45 min including time for generating a summary of the solution into a computer file. The entire team sessions were videotaped and synced to qEEG recording for each subject.

2.2 Performance Metrics

To derive performance scores, both individual and team solutions were rated by four trained coders in terms of:

- *effective problem solving* - the extent to which the case was diagnosed thoroughly and all relevant information presented in the case and expertise were utilized to solve the problem
- *decisiveness* - the extent to which one clear and explicit solution to the case was derived
- *creativity* - the extent to which a new or different and useful approach was developed that was not explicitly considered or implied in the case
- *moral reasoning* - the extent to which advanced ethical principles were used and the derived case solution showed concern for the others and the common good.

The coders worked in two teams: one team of coders rated the teams' solutions in terms of effective problem solving, decisiveness and creativity; and the other coder team rated the level of moral reasoning displayed in the solutions. The split of coding avoided any possible moral bias that may have occurred when only two coders had coded all four categories simultaneously. The coders showed high levels of agreement and inter-rater reliability in their scoring.

Furthermore, *team process* [7] can also be regarded as an outcome of teams in that teams with better task and interpersonal processes tend to perform more effectively. We assessed team process through survey ratings of respective team members at the conclusion of the team discussion. This included a combination of:

- *transition processes* - developing an overall strategy to guide the team activities
- *action processes* - ensuring that the team was using the right information to perform well
- *interpersonal process* - sharing a sense of team harmony, togetherness, and cohesion.

2.3 Psychological Metrics

Two types of psychological constructs (i.e., engagement and leadership) were obtained with multi-source psychometrics measurement procedures, both at the individual and team levels. The engagement scores relied on self-assessment, and the leadership scores involved other team member's assessment (for each respective team member) through a survey at the conclusion of the team task.

Each team member rated both his individual and overall team engagement during the task. Engagement was assessed with the scale of 14 items including physical, emotional/affective, and cognitive aspects of engagement [1].

Leadership scores for each subject were assessed by the other team members in a survey that covered the following aspects of leadership:

- *transformational leadership* [8, 9] - intellectual stimulation (i.e., helping others to examine and solve problems in new ways) and inspirational motivation (i.e., expressing confidence and enthusiasm about goals and what needed to be accomplished)
- *emergent leadership* [2, 10] - the overall degree to which the team members relied on and considered a respective team member to have shown the leadership role during the team task.

All members of a respective team rated the other members (excluding himself). As the level of agreement among the subjects was high, these scores were averaged to provide a single score for each leadership measure for each subject. These individual scores were then aggregated over all team members to attain leadership scores at the team level.

2.4 qEEG Data Recording and Signal Processing

The wireless B-Alert sensor headset [11] was used to acquire qEEG data of all subjects during the baseline 3CVT and the team discussion sessions. The qEEG recordings during the team process were synchronized with the respective videos. The qEEG data from 9 sites (POz, Fz, Cz, C3, C4, F3, F4, P3, and P4) were recorded with a sampling rate of 256 samples per second. The qEEG signals were first filtered with a band-pass filter (0.5-65Hz) before the analog to digital conversion and then the sharp notch filters were applied to remove environmental artifacts from the power network. The algorithm [11] was utilized to automatically detect and remove a number of artifacts in the time-domain qEEG signal, such as spikes caused by tapping or bumping of the sensors, amplifier saturation, or excursions that occur during the onset or recovery of saturations. Eye blinks and excessive muscle activity were identified and decontaminated by an algorithm [12] based on wavelet transformation.

From the filtered and decontaminated qEEG signal, the absolute and relative power spectral densities (PSD) were calculated on an epoch-by-epoch basis for each 1Hz bin from 1 to 40 Hz by applying fast Fourier transformation (FFT) to the 50% overlapping 1sec overlays of the qEEG data. In order to reduce the edge effect, the Kaiser window was applied to each overlay. Furthermore, the FFT on three successive overlays was averaged to decrease epoch-by-epoch variability. The following PSD bandwidths were extracted: theta slow, theta fast, theta total, alpha slow, alpha fast, alpha total, beta, and gamma.

In order to explore the applicability of neurological alertness quantification in estimation of the psychological metrics, we also included into the analysis the outputs of the B-Alert model [12, 13] that quantifies engagement levels and identifies cognitive state changes. It is an individualized model that selects the most discriminative PSD variables, derives coefficients for a discriminant function, and classifies subject's cognitive state for each epoch into one of the four levels of alertness (sleep onset, distraction/relaxed wakefulness, low engagement, and high engagement).

As we are dealing with the high-level psychological constructs, both types of epoch-by-epoch variables (i.e., PSD bandwidths and B-Alert classification probabilities) were then averaged over a 30sec sliding window in 1sec increments to get a general trend of neurological changes over time.

To normalize the qEEG data for individual variability, the absolute and relative PSD values during the teaming task were z-scored to the qEEG data during the baseline 3CVT session for each respective subject. Similarly, the B-Alert output engagement probabilities during the team discussion were also calibrated in the same manner by z-scoring them with respect to the subject's engagement during the 3CVT task.

2.5 Data Analysis

First, correlation analysis was performed to explore if psychological metrics such as engagement and leadership relate to the achieved team performance scores in the conducted teaming study.

Second, different modeling approaches were applied to estimate the psychological metrics based on the qEEG data. Both individual and team scores of engagement and leadership were analyzed and grouped into two classes in the following manner: all team scores that were above the overall mean value for all teams were considered as "High", and all team scores that were below the mean value were considered as "Low". In order to accommodate potential differences among different teams, individual scores were first z-scored with respect to the mean value within the respective team. Next, such normalized individual scores that were above 0 (i.e. above average value for the respective team) were grouped into the "High" class, and the ones that were below 0 were assigned to the "Low" class. The most discriminative qEEG variables were selected by step-wise variable selection procedure. Afterwards, the selected variables were used in three different algorithms: linear DFA, quadratic DFA, and linear regression. Linear and quadratic DFA classify the data into the two classes of interest (i.e., "High" and "Low"). Linear regression algorithm predicts the value of psychological variable and then we classify it into one of the two classes based on the above defined thresholds for the classes.

Table 1. Statistically significant correlations between psychological measures and performance variables at the individual and team levels

Individual scores			
Psychological measure	Performance	Correlation	p
Engagement	Decisiveness	0.16	0.03
Transformational leadership	Moral reasoning	0.14	0.05
Emergent leadership	Moral reasoning	0.21	0.01
Team scores			
Engagement	Team process	0.58	0.0002
Transformational leadership	Team process	0.34	0.03
Emergent leadership	Moral reasoning	0.31	0.04

As the goal is to recognize the team members who are highly engaged and/or good leaders, the "High" class is assumed to be the positive class and the trained algorithms were evaluated in terms of the models' sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) in two different ways: auto-validation

and cross-validation. In the initial model development phase, auto-validation is crucial to test the feasibility of the model by testing the model on the training data. Cross-validation assesses the generalization capabilities of the model by testing it on the data that was not used for training. In order to examine team independent training, we performed leave-one-team-out cross-validation by training the model on the data from all teams but one, and testing the trained model on the removed team's data. The procedure was repeated for all teams in the study, and the results were averaged across all cross-validation rounds.

3 Results

In this section the following results are presented: (1) statistically significant correlations between the psychological measures of engagement and leadership at the individual and team levels and the corresponding performance scores, (2) classification results of the applied algorithms (linear DFA, quadratic DFA, and linear regression), and (3) detailed analysis of different experimental settings.

3.1 Correlations

Correlation analysis showed that psychological measures, such as engagement and leadership are positively correlated with some of the achieved performance scores both at the individual and team levels. In Table 1 are shown statistically significant correlations (based on the two-tailed t-test). From the table one can observe that engagement is related to the team process and decisiveness, while leadership is correlated with the team process and moral reasoning. On the other hand, performance scores such as effective problem solving and creativity were not correlated with the analyzed metrics of engagement and leadership.

3.2 Classification Results

The auto-validation and leave-one-team-out cross-validation classification results of the three evaluated algorithms (i.e., linear DFA, quadratic DFA, and linear regression) for engagement, transformational and emergent leadership at the individual and team levels are shown in Table 2 and Table 3, respectively. The results are averaged over 30 teams, as the qEEG data for one team were unusable due to difficulties in capturing qEEG data and synchronization problems between video and qEEG recordings.

Typically, quadratic DFA achieved the highest auto-validation classification results. However, leave-one-team-out cross-validation showed that the quadratic DFA classifier achieved significantly higher specificity than sensitivity. After investigating the confusion matrices in such cases, it turned out that majority of the instances were classified as "Low" class in such cases.

Table 2. Classification results for the psychological measures at the individual level. Three algorithms were evaluated: linear discriminant function analysis (L-DFA), quadratic discriminant function analysis (Q-DFA), and linear regression through auto-validation (AV) and leave-one-team-out cross-validation (CV).

Individual level measures								
Engagement								
	Sensitivity		Specificity		PPV		NPV	
Algorithm	AV	CV	AV	CV	AV	CV	AV	CV
L-DFA	80.8%	47.0%	79.4%	48.7%	78.4%	45.8%	81.8%	49.9%
Q-DFA	72.8%	15.5%	99.6%	80.2%	99.4%	41.9%	79.9%	50.7%
Regression	69.1%	36.8%	79.3%	51.1%	75.5%	41.5%	73.6%	47.2%
Transformational leadership								
	Sensitivity		Specificity		PPV		NPV	
Algorithm	AV	CV	AV	CV	AV	CV	AV	CV
L-DFA	75.6%	40.9%	76.3%	49.9%	76.5%	45.5%	75.4%	45.3%
Q-DFA	74.8%	10.2%	99.4%	79.1%	99.2%	33.2%	79.4%	46.3%
Regression	56.8%	32.3%	83.1%	57.6%	77.4%	43.7%	65.5%	45.4%
Emergent leadership								
	Sensitivity		Specificity		PPV		NPV	
Algorithm	AV	CV	AV	CV	AV	CV	AV	CV
L-DFA	75.2%	40.0%	75.8%	46.0%	75.9%	42.9%	75.0%	43.0%
Q-DFA	77.7%	18.5%	99.3%	75.6%	99.2%	43.5%	81.4%	47.7%
Regression	58.6%	31.0%	81.0%	52.5%	75.8%	39.8%	65.8%	42.8%

Overall, linear DFA performed well and obtained acceptable results in all settings. Even though, in some cases the cross-validation results were below the chance level, some metrics such as emergent leadership at the team level were accurately recognized. As this seem to be the most promising classifier, in the next section we further analyze its performance in different settings.

3.3 Analysis of Different Experimental Settings

Next, we analyze the effects of the qEEG data normalization with respect to the baseline 3CVT session and averaging over the sliding window. The classification results of the linear DFA classifier for emergent leadership at the team level are shown in Table 4 for four different settings:

- both 3CVT normalization and sliding window are applied
- only 3CVT normalization is applied
- only sliding window is applied
- neither 3CVT nor sliding window is applied

Table 3. Classification results for the psychological measures at the team level. Three algorithms were evaluated: linear discriminant function analysis (L-DFA), quadratic discriminant function analysis (Q-DFA), and linear regression through auto-validation (AV) and leave-one-team-out cross-validation (CV).

Team level measures								
Engagement								
	Sensitivity		Specificity		PPV		NPV	
Algorithm	AV	CV	AV	CV	AV	CV	AV	CV
L-DFA	77.0%	53.5%	77.4%	45.4%	71.0%	41.4%	82.3%	57.5%
Q-DFA	95.3%	28.8%	98.4%	66.2%	97.8%	38.1%	96.6%	56.3%
Regression	64.5%	45.8%	70.9%	56.4%	61.5%	43.1%	73.4%	59.1%
Transformational leadership								
	Sensitivity		Specificity		PPV		NPV	
Algorithm	AV	CV	AV	CV	AV	CV	AV	CV
L-DFA	81.5%	53.9%	77.5%	46.2%	74.4%	44.5%	84.0%	55.7%
Q-DFA	99.5%	68.6%	84.5%	39.0%	83.7%	47.3%	99.6%	60.8%
Regression	79.2%	59.6%	64.2%	42.2%	63.9%	45.2%	79.4%	56.7%
Emergent leadership								
	Sensitivity		Specificity		PPV		NPV	
Algorithm	AV	CV	AV	CV	AV	CV	AV	CV
L-DFA	82.2%	64.2%	81.4%	55.1%	80.6%	57.3%	83.0%	62.2%
Q-DFA	99.5%	73.0%	87.3%	34.2%	88.1%	51.0%	99.4%	57.4%
Regression	79.6%	64.9%	69.7%	48.3%	71.1%	54.1%	78.4%	59.5%

From the table, it can be clearly seen that the classifier benefits from both normalization and averaging of data. On average, when both normalization and averaging were applied the classification results were higher by 24.6% and 13.2% in the case of auto-validation and cross-validation, respectively. When looking at the improvements that normalization and averaging bring separately, it turned out that averaging is more valuable as the results were slightly more improved that way.

4 Discussion

The current study aimed at developing a method for estimation of psychological measures in the team setting based on the neurophysiological data. In order to achieve that goal, the teams were rated in terms of engagement and leadership while solving a corporate social responsibility case. The qEEG data were utilized during the team discussion to provide insight into brain activity of the team members. The objective was to meet the three criteria: (1) The algorithm had to be computationally simple so that it could be easily implemented in real-world applications; (2) The approach had to accommodate individual variability in the qEEG data; (3) The approach had to capture a general trend of the qEEG changes over time in order to address high-level psychological constructs. Next, we summarize how each of these criteria was met, and discuss the limitations of the algorithm and future work directions.

Table 4. Comparison of linear DFA classification results for emergent leadership at the team level for different 3CVT normalization and sliding window settings

Setting		Sensitivity		Specificity		PPV		NPV	
3CVT normalization	sliding window	AV	CV	AV	CV	AV	CV	AV	CV
yes	yes	82.2%	64.2%	81.4%	55.1%	80.6%	57.3%	83.0%	62.2%
yes	no	68.3%	60.4%	61.7%	51.9%	62.7%	54.2%	67.3%	58.2%
No	yes	66.8%	45.5%	68.4%	48.5%	62.8%	41.4%	72.0%	52.7%
No	no	56.0%	44.9%	58.5%	47.8%	52.0%	40.9%	62.3%	51.9%

Three algorithms were evaluated: linear DFA, quadratic DFA, and linear regression. These algorithms were chosen as they are relatively simple, but still very effective in different application scenarios. Based on the experimental results, in our study, linear DFA proved to be the most effective in predicting both psychological metrics of interest, i.e. engagement and leadership. The algorithm successfully coped with individual qEEG data variability by calibrating the data with respect to the baseline sustained attention task (i.e., 3CVT). The epoch-by-epoch data variability was reduced by averaging the data over a 30sec sliding window. Both data normalization and averaging substantially improved the classification results.

The proposed algorithm demonstrated the feasibility of neurophysiological estimation of team psychological metrics. One of the main findings of our work is that the qEEG data carry a wealth of information and can help assessing different aspects of team process. This is only a first step towards a broader acceptance of the qEEG-based psychological assessment. In order to move beyond controlled laboratory experiments, a few limitations of our work need to be overcome. First, the psychological measures were assessed by either the subjects themselves or other team members. Such scores might be biased and slightly subjective. We are re-assessing the scores by the trained coders. However, as shown in Section 3.1 our psychological measures were correlated with the objective team performance scores. Second, the cross-validation results were noticeably lower than the auto-validation results. That could be altered by team-dependent training which would require longer recordings of the team process to acquire sufficient amount of training data to model the complex psychological constructs such as engagement and leadership. Third, parts of the qEEG recordings were unusable due to the large amount of noise (i.e., artifacts) in the data. In order to enable unobtrusive long-term qEEG recordings in realistic settings, we are streamlining our platform, especially in context of the team settings, by further improving the acquisition system, timing accuracy, and the artifact decontamination algorithms. Fourth, the study focused on business students who were solving particular case of social responsibility. In the future, we plan to extend the study to real-world organizations with specific tasks and team roles. Lastly, we will examine in more details which brain regions might be indicative of engagement and leadership.

References

1. Rich, B.L., Lepine, J.A., Crawford, E.R.: Job Engagement: Antecedents and Effects of Job Performance. Academy of Management 53, 617–635 (2010)
2. Carson, J.B., Tesluk, P.E., Marrone, J.A.: Shared Leadership in Teams: An Investigation of Antecedent Conditions and Performance. Academy of Management Journal 50, 1217–1234 (2007)
3. Waldman, D.A., Balthazard, P.A., Peterson, S.J.: Social Cognitive Neuroscience and Leadership. The Leadership Quarterly 22, 1092–1106 (2011)
4. Stevens, R., Galloway, T., Wang, P., Berka, C., Tan, V., Wohlgemuth, T., Lamb, J., Buckles, R.: Modeling the Neurodynamic Complexity of Submarine Navigation Teams. Computational and Mathematical Organization Theory (2012)
5. Balthazard, P.A., Waldman, D.A., Thatcher, R.W., Hannah, S.T.: Differentiating Transformational and Non-transformational Leaders on the Basis of Neurological Imaging. The Leadership Quarterly 23, 244–258 (2012)
6. Pless, N., Maak, T.: Levi Strauss & Co: Addressing Child Labour in Bangladesh. In: Mendenhall, M.E., Oddou, G.R., Stahl, G.K. (eds.) Readings and Cases in International Human Resource Management and Organizational Behavior, 5th edn., pp. 446–459. Routledge, London (2011)
7. Marks, M.A., Mathieu, J.E., Zaccaro, S.J.: A Temporally Based Framework and Taxonomy of Team Processes. Academy of Management Review 26, 356–376 (2001)
8. Balthazard, P.A., Waldman, D.A., Warren, J.E.: Predictors of the Emergence of Transformational Leadership in Virtual Decision Teams. The Leadership Quarterly 20, 651–663 (2009)
9. Bass, B.M., Avolio, B.J.: The Multifactor Leadership Questionnaire. Consulting Psychologists Press, Palo Alto (1990)
10. Zhang, Z., Waldman, D.A., Wang, Z.: A Multilevel Investigation of Leader-Member Exchange, Informal Leader Emergence, and Individual and Team Performance. Personnel Psychology 65, 49–78 (2012)
11. Berka, C., Levendowski, D.J., Cvetinovic, M.M., Petrovic, M.M., Davis, G., Lumicao, M.N., Zivkovic, V.T., Popovic, M.V., Olmstead, R.: Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset. International Journal of Human-Computer Interaction 17, 151–170 (2004)
12. Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L.: EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. Aviation Space and Environmental Medicine 78, B231–B244 (2007)
13. Johnson, R.R., Popovic, D.P., Olmstead, R.E., Stikic, M., Levendowski, D.J., Berka, C.: Drowsiness/Alertness Algorithm Development and Validation Using Synchronized EEG and Cognitive Performance to Individualize a Generalized Model. Biological Psychology 87, 241–250 (2011)

Physio-behavioral Coupling as an Index of Team Processes and Performance: Overview, Measurement, and Empirical Application

Adam J. Strang[1], Gregory J. Funke[2], Sheldon M. Russell[2], and Robin D. Thomas[3]

[1] Consortium Research Fellows Program, Alexandria VA, USA
[2] Air Force Research Laboratory, Wright-Patterson AFB OH, USA
[3] Miami University, Oxford OH, USA
adam.strang.ctr@us.af.mil

Abstract. Research shows that teammates often exhibit similarity in their physiological and behavioral responses during cooperative task performance, a phenomenon referred to here as physio-behavioral coupling (PBC). Goals of this manuscript are to provide an overview of research examining the utility of PBC as an index of team processes (e.g., coordination) and performance, discuss applied and theoretical issues in PBC measurement, and present findings from a study using linear and nonlinear statistics to assess PBC.

Keywords: Team, Coupling, Coordination, Performance, Nonlinear.

1 Introduction

This manuscript supports the parallel session entitled, "Modeling the Complex Dynamics of Teamwork." The focus of this manuscript is physio-behavioral coupling (PBC), its relation to team processes and performance, and issues regarding its measurement and interpretation, with an emphasis on practical applications. An experimental illustration of these issues is provided.

1.1 PBC, Team Processes, and Team Performance

PBC can be defined as a statistical similarity in the cortical, autonomic, or behavioral activity of two or more members of a team engaged in cooperative behavior. Over the past three decades, researchers have identified PBC in a number of physio-behavioral responses, including cardiac inter-beat intervals (IBIs), electrical brain activity, and human postural sway, and in diverse team task environments, such as military room clearing, team puzzle solving, and duet guitar playing [1-5]. In many cases, PBC manifests as an emergent (i.e., spontaneous and self-organized) phenomenon outside conscious awareness [1]. As such, there has been speculation about the underlying causes (or drivers) of PBC, its role in cooperative task performance, and the associations it shares with important team processes (e.g., strategy, coordination, communication, cohesion, etc.).

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 219–228, 2013.

Most explanations posit that PBC reflects important team-level processes, such as communication and coordination [1]. For example, oral communication is a vector employed by teams to discuss strategy and coordinate action [6]. Research has demonstrated that oral communication is sufficient to drive the coupling of human postural sway, supporting speculation that sway coupling can serve as an indirect index of team communication dynamics [4] – a speculation that our own research supports [5]. Other research has identified PBC between group members performing very different actions (e.g., active participants and passive observers) [7], supporting the perspective that PBC may be caused by emotional (arousal) and/or cognitive (shared situation awareness) dynamics associated with group/team membership. Finally, PBC has been shown to exhibit relationships with psychosocial phenomenon like rapport and trust [8, 9], prompting some to speculate that PBC may facilitate, rather than simply reflect a consequence of, team processes.

To date, only a small number of studies have examined the association between PBC and team performance. While several studies suggest a positive relationship (i.e., higher PBC is related to better team performance) [2, 3], our research indicates that a negative relationship is possible [5].

From an applied perspective, PBC has been shown to exhibit a moderate relationship with performance (absolute $r \sim .4$) [1, 5], which is comparable in magnitude to correlations observed between performance and other team processes such as cohesion and collective efficacy (both $r \sim .25$) [10, 11]. This suggests that PBC measures may have an advantage over other (largely self-reported) team process assessments since many responses used to estimate PBC are minimally invasive (e.g., cardiac IBI, postural sway), and many metrics of PBC can be computed in real-time without interrupting task performance.

1.2 PBC Measurement: A Historical Review and Recent Developments

PBC has been characterized using a variety of different statistical measures, leading to inconsistency across studies. It has also been exceedingly rare for researchers to communicate why a particular measure (or set of measures) was chosen to over others (see [7] and [12] for exceptions). However, choosing the proper measure is critical since it may influence the ability to detect meaningful changes in PBC, determine the information about PBC obtained (e.g., coupling strength versus phase relation), and have implications for the utility of PBC measurement in applied applications.

Early PBC studies used independent rater analysis of recorded video and/or physiological signals to detect response similarities [8]. While these methods were carefully implemented, they are subjective, as well as both cost and time prohibitive.

In more recent studies assessing PBC in cyclical motor tasks (e.g., swinging of handheld pendulums), researchers have often employed relative phase statistics [13]. Though relative phase is an intuitive indicator of synchronicity (a specific sub-type of coupling) and phase relation (e.g., in-phase versus out-of-phase), it is effective primarily for examining responses that exhibits near-sinusoidal oscillations [14].

An additional approach employed by Henning and colleagues [2] has been to examine PBC using cross-correlation (CC) and cross-spectral coherence (CSC), which

are linear statistics that describe the degree of similarity between two time series in the time and frequency domains, respectively. Advantages of the measures include a long and accepted history for examining complex time-series data and the ability to provide estimates of multiple coupling dynamics (CC: coupling strength and temporal lag; CSC: coupling strength at particular frequencies) in near-real-time [15]. Disadvantages include linear assumptions of periodicity and stationarity (i.e., equal mean and variance), which many physio-behavioral responses are known to violate [16, 17].

In an attempt to overcome the limitations of linear statistics, some researchers (including the current authors) have explored the use of nonlinear measures to characterize PBC. Although computationally quite different from one another, this family of statistics, which include measures such as Cross-Recurrence Quantification Analysis (CRQA) [12], Cross Sample Entropy (CSEn) [18], and Average Mutual Information (AMI) [19], can be used to confirm the existence of nonlinear coupling and quantify its strength. To illustrate how the information obtained from linear and nonlinear coupling measures differ, consider that CC, when a zero lag is employed, is equivalent to a Pearson product-moment correlation [20]. Thus, CC characterizes the degree to which two time-series share a one-to-one (i.e., linearly synchronized) relationship in both time and (relative) amplitude. Conversely, nonlinear measures (with acknowledgment that the following is a broad generalization) quantify the degree to which two time series exhibit matching temporal patterns (i.e., strings of sequential data points) across an entire time interval, regardless of where those matches occur within that interval. Thus, nonlinear measures do not index synchronicity (a potential limitation if this is the coupling dynamic of specific interest), but rather the overall degree of patterning shared between two data streams within a specified temporal envelope.

It is because of this flexibility that nonlinear measures may be better suited for detecting and quantifying coupling strength in aperiodic and noisy systems [12]. This view is supported by findings that nonlinear coupling measures demonstrate greater sensitivity, compared to more traditional linear measures like CC and CSC, for detecting changes in coupling dynamics among paired physical systems [21], financial trends [22], human postural sway [12], and animal neurophysiological responses [23].

However, claiming that nonlinear coupling measures are more sensitive than linear measures, without first identifying that the systems under examination exhibit meaningful (i.e., deterministic) linear and/or nonlinear coupling, is problematic. To establish that meaningful coupling is evident, surrogation tests are required.

The most straightforward and intuitive method to perform surrogation tests first involves obtaining estimates of PBC (for each metric of interest) from originally sampled time-series representing the response(s) of interest (e.g., postural sway from two people engaged in oral communication). Next, new (surrogate) time-series are generated by (separately) randomly shuffling the sequence of data points within each original time-series. The result of this procedure are two time-series in which any deterministic temporal structure that originally existed in individual responses, as well as any meaningful coupling between those responses, is eliminated. Then, PBC estimates are obtained for the surrogate time-series and compared with those from the original time-series. If the two sets of PBC estimates are shown to be equivalent (often determined using inferential statistics applied to an entire experiment's sample), this suggests that no meaningful coupling existed in the original time-series. Conversely, if PBC estimates in the original time-series are greater than those observed

from the surrogates, this suggests that meaningful coupling does exist in the original. In a case where both linear and nonlinear measures are used to examine PBC, and both detect meaningful coupling, then it is possible to examine the PBC metrics for sensitivity differences, with the understood caveat that each type of metric characterizes a different coupling dynamic. However, in a case where meaningful nonlinear coupling is detected but linear coupling is not, investigating sensitivity differences is futile, since the very application of linear statistics in this case is inappropriate.

To date, very few PBC studies have included any form of surrogation test. However, it is our view that these tests are critical since they not only provide useful information about the underlying dynamics of a coupled relationship, but also verify the appropriateness of statistics used to draw inference about the phenomenon.

2 Empirical Application of Linear and Nonlinear Measures for Assessing PBC in a Cooperative Team Task

The remainder of this manuscript is dedicated to describing methods and results from a single experiment in which PBC was examined in dyads performing a cooperative pointing task (Fig. 1). The purpose of including this experiment here is to provide guidance on application of linear and nonlinear measures to examine PBC, as well as explore the unique (or analogous) information linear and nonlinear metrics may provide about team coordination and performance.

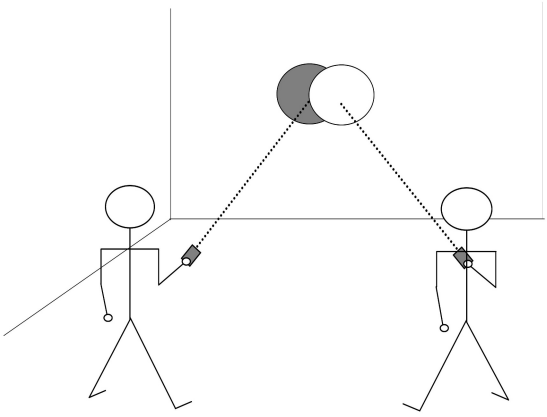


Fig. 1. Illustration of the cooperative pointing task used in this experiment. The goal of the task was to achieve complete overlap of two "virtual" circles projected onto a far wall by manipulating handheld remote controls (Nintendo Wii Remotes) that controlled the circles' movements.

The *pointing task* employed in this experiment required dyads to control the movements of two "virtual" circles (each participant controlled one circle) and align them such that the two circles completely overlapped one another. This particular task was selected because it presented a context in which PBC (i.e., wrist coupling), team coordination, and team performance should be closely related. Thus, we hoped in

utilizing this task we might draw more direct inference from PBC measurement about the coordination dynamics supporting team performance, and about how those dynamics are altered by a manipulation of task constraints (i.e., an increase in difficulty). Other goals included a determination of the presence of meaningful linear and/or nonlinear coupling in team members' wrist movements (using surrogation tests) and a comparison of the sensitivity of linear and nonlinear coupling measures to team performance differences across task difficulty conditions.

2.1 Methods

To perform the pointing task, dyads (N = 30 pairs) stood upright facing a far wall at a distance of 1.5 m while holding handheld remote controls (Nintendo Wii remotes – "wiimotes") in their dominant hand with elbow flexed to 90 degrees. Two large circles (145 and 150 pixels in diameter, respectively) were shown on a far wall using a video projector. The vertical and lateral movements of the circles were linked to the movements of the wiimotes using custom software and Bluetooth connection. The task performance goal was to achieve complete overlap of the two circles (at a location near the center of the wall) and hold that position for "as long as possible" in 90 second trials.

Task difficulty was manipulated by altering the wiimote-to-circle movement ratio. In the *easy* condition the ratio was 1:.25, meaning that a 1 cm translation of the wiimote elicited a .25 cm translation of the circle to which it corresponded. In the *normal* and *hard* conditions the ratio was 1:1 and 3:1, respectively.

Dyads performed two trials of each condition in counterbalance order (six trials total). Throughout trials wrist movements (yaw and pitch rotation) were recorded from participants at 75 Hz using two wireless Xsens Technologies Mtw inertial trackers. In post-processing, yaw and pitch time-series were cropped to 60 second durations by removing the first and last 15 seconds of each trial. The truncated times-series were then subjected to .1 to 30 Hz 2nd order Band-pass Butterworth filters (to achieve stationarity and eliminate high frequency noise) and normalized to unit variance.

Normalized time-series were then paired within dyad and rotational plane and examined using CC, CRQA (percent recurrence; %REC)[1], CSEn[2] and AMI in 13.65 sec (1,024 data point) windows with a 6.83 (512 point) overlap.[3] This procedure rendered six values in each trial for each PBC measure; from those six values, the median was recorded for each measure to indicate the central tendency of yaw and pitch wrist coupling. Median PBC estimates from like conditions were then averaged.

---

[1] %REC is the percentage of points (where point represents a distance vector comprised of a serial sequences of data values) that repeat in a 2-dimenstional recurrence plot. It serves as an indicant of the overall amount of patterning in a time-series [12].

[2] In subsequent reporting the inverse of CSEn, $CSEn^{-1}$, is presented to facilitate directional correspondence with interpretation of all other PBC measures.

[3] CC was estimated with zero lag, replicating the procedures of [2]. CSEn parameters, M = 3 (vector lengths for comparison) and r = .3 (vector tolerance) were set using a parameter selection procedure described by [24]. CRQA parameters, i.e., embedding dimension (*EmD* = 9), time delay (*td* = 4), rescaling method (*rescale* = euclidean), and radius (*rad* = 10), were established using the procedure described by [12]. AMI requires no parameter selection using the algorithm provided in [19].

2.2 Results

Surrogation tests were used to determine whether meaningful linear and/or nonlinear coupling existed between wrist movements of dyads using identical methods to those described earlier in section 1.2. Inferential comparisons testing for differences in PBC estimate from original and surrogate time-series were carried out using paired samples *t*-tests for all PBC metrics.

Results indicated that, across nonlinear measures, PBC estimates from the original time-series were significantly greater than those obtained from surrogates. However, no difference was found between PBC estimates of original and surrogate time-series for CC (Table 1). This indicates that the wrist movements of teammates exhibited a nonlinear, as opposed to linear, coupled relationship. From a practical perspective this means that dyad wrist movements did not exhibit linear synchronicity, though they did exhibit meaningful similarities in overall temporal patterning. Consequently, CC was dropped from further analyses and comparisons.

Table 1. Mean of median PBC estimates and standard errors (in parentheses) obtained from original and surrogate (randomly shuffled) time-series pairings

PBC Measure	Original pairs	Surrogate pairs	*t*
CC – yaw	-.01 (.01)	.00 (.00)	.53
CC – pitch	-.03 (.01)	.00 (.00)	1.87
CSEn[1] - yaw	11.01 (.25)	.68 (.00)	41.40*
CSEn[1] - pitch	10.76 (.33)	.67 (.00)	30.14*
%REC – yaw	3.95 (.13)	.00 (.00)	30.39*
%REC - pitch	3.83 (.13)	.00 (.00)	29.71*
AMI - yaw	.29 (.00)	.03 (.00)	56.64*
AMI - pitch	.30 (.05)	.03 (.00)	55.32*

Note. *t*-crit $_{df = 89,\ \alpha = .05}$ = 1.99.

* $p < .05$

Next, effects of *task difficulty* were examined using separate repeated-measures ANOVAs for PBC measures and the team performance metric, Circle Overlap[4]. Omnibus main effects of Circle Overlap, F (1.60, 46.31) = 2701.14, $p < .05$, and AMI-yaw, F (1.87, 54.33) = 6.89, $p < .05$, were detected. Post-hoc pairwise comparisons revealed a precipitous decline in Circle Overlap as a product of increases in *task difficulty* (Fig. 2a). Circle Overlap decreased as task difficulty increased, which confirms that the experimental manipulation was effective in diminishing team performance.

Post-hoc assessment of AMI-yaw revealed lower wrist coupling in the *normal* and *hard*, as compared to the *easy* condition – indicating that nonlinear wrist coupling decreased as a result of increases in task difficulty (Fig. 2b).

[4] Circle Overlap is defined as the cumulative time (in seconds) during a 90 second trial that complete overlap of the two circles was achieved.

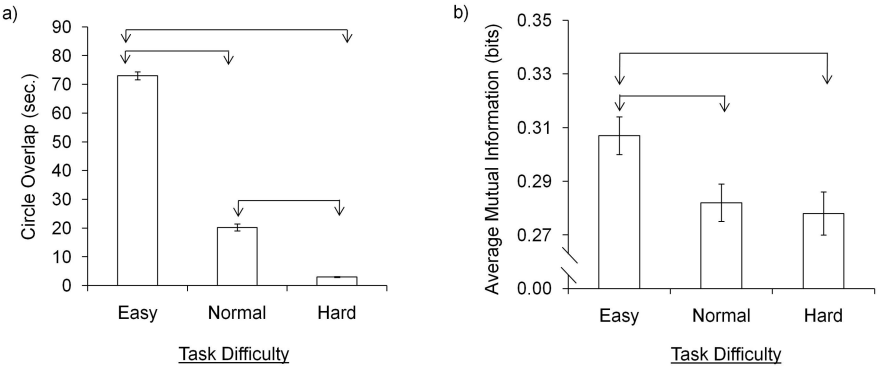


Fig. 2. Mean estimates of Circle Overlap (a) and Average Mutual Information in wrist-yaw rotation (b) across task difficulty conditions. Connected arrows represent significant pair-wise differences at $p < .05$. Error bars are standard errors.

Finally, Pearson (r) correlations were used to assess the relationship between team performance (i.e., Circle Overlap) and nonlinear wrist coupling strength for each *task difficulty* condition (Table 2). Interestingly, all of the nonlinear metrics we employed were analogous in detecting a negative relationship between PBC and performance across all conditions, indicating that a decrease in coupling was related to better performance. In other words, lower similarity in the temporal patterning of coupled wrist movements was related to a greater ability to keep the two circles aligned. The similar direction of effect detected across measures is also intriguing, and suggests that these measure, though computational quite different form one another, are sensitive to similar dynamical properties about a coupled relationship (as argued in section 1.2).

Still, there were noticeable differences in the magnitude of relationship detected between wrist coupling and team performance across measures. Specifically, %REC detected the strongest relationship (r = -.50, collapsed across condition and rotational plane), followed by $CSEn^{-1}$ (r = -.40) and AMI (r = -.26), respectively. This may indicate sensitivity differences between these measures, at least in-so-far as accounting for individual differences in team performance are concerned.

Table 2. Pearson correlations between PBC measures and team performance (Circle Overlap) for each task difficulty condition

	Task Difficulty Condition		
PBC Measure	Easy	Normal	Hard
$CSEn^{1}$ - yaw	-.38*	-.54*	-.25
$CSEn^{1}$ - pitch	-.48*	-.41*	-.33
%REC – yaw	-.51*	-.31	-.56*
%REC - pitch	-.41*	-.68*	-.50*
AMI - yaw	-.43*	-.40*	-.23
AMI - pitch	-.32	-.13	-.03

Note. r-crit $_{df = 28,\ \alpha = .05}$ = .36

* $p < .05$

2.3 Discussion

Dyad wrist movements did not exhibit a linear, but rather a nonlinear coupled relationship in this experiment. This is important because it determined the family of statistics that were appropriate for examining PBC experimental effects, but also because it ruled out synchronicity (the coupling dynamic indexed by CC) as a coordination strategy that could have been utilized by teams to perform the task. However, we make this statement with some caution since our results do not exclude the possibility that a synchronized relationship might have existed at some temporal lag, potentially indicating a linear leader-follower coordination strategy. We are currently examining this issue.

Findings regarding the manipulated effect of task difficulty revealed lower nonlinear wrist coupling strength (reduced AMI) in conditions where the task was more difficult. In interpreting this effect, it is first important to mention that this finding does not insinuate that wrist movements were completely decoupled, since significant nonlinear coupling was confirmed through surrogation tests. Rather, this finding indicates that nonlinear coupling strength was simply *less* in higher difficulty conditions. Second, if the only other information provided is that performance was also decreased by increases in task difficulty, then one possible explanation is that higher difficulty may have inhibited the ability of teams to coordinate effectively, leading to decreased coupling and reduced performance. However, insight gained from the correlation analyses supports a different interpretation; namely, that a decrease in wrist coupling under higher task difficulty may have reflected a compensatory strategy. This interpretation is supported by the ubiquitous set of negative correlations detected between nonlinear wrist coupling and team performance across task difficulty conditions, indicating that decreases in wrist coupling strength were associated with increases in performance.

As mentioned in section 1.1, we found a similar relationship between team performance and PBC (in cardiac IBIs) in a previous study [5]. In that study, we posited that a negative correlation may have indicated general team coordination plasticity or a complimentary coordination strategy featuring asynchronous and/or anti-phase team member behaviors [25] – either of which could result in decreases in nonlinear coupling strength. Here we come to similar conclusions.

In considering sensitivity differences in the set of nonlinear measures we employed, it appears that our results lead to mixed interpretations. On one hand, because AMI was the only measure to detect meaningful changes in wrist coupling induced by the experimental manipulation of task difficulty, it could be argued that this metric was more sensitive than the others. However, %REC exhibited the strongest correlations with performance, hinting that it was best in accounting for individual team performance differences. While the results of this study cannot definitively address issues of measure sensitivity across coupling metrics, they raise interesting possibilities. To formally address the issue further we have planned a series of modeling experiment wherein coupling strength will be mathematically manipulated, allowing for true quantitative comparisons.

3 Conclusion

In this manuscript we presented an overview of PBC with a focus on studies that have explored its utility as an index of team processes and performance. In addition, we provided an overview of important measurement issues in PBC research, followed by a simple empirical study that contextualized and accentuated this matter. Overall, we believe that examination of PBC is a fruitful area for ongoing research, since it not only appears to be informative for basic theory development of team dynamics but also has the potential for use in real-time team monitoring.

Acknowledgements. This research was generously supported by an Air Force Office of Scientific Research (AFOSR) grant (Program Manager: Dr. Jay Myung).

References

1. Knoblich, G., Butterfill, S., Sebanz, N.: Psychological research on joint action: Theory and data. In: Ross, B. (ed.) The Psychology of Learning and Motivation. Academic Press, Burlington (2011)
2. Henning, R.A., Boucsein, W., Gil, M.C.: Social-physiological compliance as a determinant of team performance. Int. J. Psychophysiol. 40, 221–232 (2001)
3. Elkins, A.N., Muth, E.R., Hoover, A.W., Walker, A.D., Carpenter, T.L., Switzer, F.S.: Physiological compliance and team performance. Appl. Ergon. 40, 997–1000 (2009)
4. Shockley, K., Baker, A.A., Richardson, M.J., Fowler, C.A.: Articulatory constraints on interpersonal postural coordination. J. Exp. Psychol. Human. 33, 201–208 (2007)
5. Strang, A.J., Funke, G.J., Knott, B.A., Warm, J.S.: Physio-behavioral synchronicity as an index of processes supporting team performance. Human Fac. Erg. Soc. P. 55, 1447–1451 (2011)
6. Salas, E., Fiore, S.M.: Team cognition:Understanding the factors that drive process and performance. American Psychological Association, Washington, DC (2004)
7. Konvalinka, I., Xygalatas, D., Bilbulia, J., Schjodt, U., Jegindo, E., Wallot, S., Van Orden, G., Roepstorff, A.: Synchronized arousal between performers and related spectators in a fire-walking ritual. P. Natl. A. Sci. USA 108, 8514–8519 (2011)
8. Bernieri, F.J.: Coordinated movement and rapport in teacher-student interactions. J. Nonverbal Behav. 12, 120–138 (1988)
9. Wiltermuth, S.S., Heath, C.: Synchrony and cooperation. Psycholog. Sci. 20, 1–5 (2009)
10. Mullen, B., Copper, C.: The relation between group cohesiveness and performance: integration. Psychol. Bull. 115, 201–227 (1994)
11. Gully, S.M., Incalcaterra, K., Joshi, A., Beaubien, J.M.: A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. J. Appl. Psychol. 87, 819–832 (2002)
12. Shockley, K.: Cross recurrence quantification of interpersonal postural activity. In: Riley, M.A., Van Order, G.C. (eds.) Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences, `http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp` (accessed February 1, 2013)
13. Schmidt, R.C., O'Brien, B.: Evaluating the dynamics of unintended interpersonal coordination. Ecol. Psychol. 9, 189–206 (1997)

14. Peters, B.T., Haddad, J.M., Heiderscheit, B.C., Van Emmerick, R.E.A., Hamill, J.: Limitations in the use and interpretation of continuous relative phase. J. Biomech. 36, 271–274 (2003)
15. Chatfield, C.: The analysis of time series: An introduction, 6th edn. CRC Press, New York (2004)
16. Carroll, J.P., Freedman, W.: Nonstationary properties of postural sway. J. Biomech. 26, 409–416 (1993)
17. Zbilut, J.P., Webber, C.L., Zak, M.: Quantification of heart rate variability using methods derived from nonlinear dynamics. In: Drzewiecki, G.M., Li, J.K. (eds.) Analysis and Assessment of Cardiovascular Function, pp. 324–334. Springer, New York (1998)
18. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate and sample entropy. Am. J. Physiol-Heart C. 278, H2039–H2049 (2000)
19. Thomas, R.D., Moses, N.C., Semple, E.A., Strang, A.J.: A usable and efficient algorithm for the computation of average mutual information: Validation and implementation in Matlab. J. Math. Psycholog. (under review)
20. Pereda, E., Quiroga, R.Q., Bhattacharya, J.: Nonlinear multivariate analysis of neurophysiological signals. Prog. Neurobiol. 77, 1–37 (2005)
21. Shockley, K., Butwill, M., Zbilut, J.P., Webber, C.L.: Cross recurrence quantification of coupled oscillators. Phys. Lett. A 305, 59–69 (2002)
22. Liu, L.Z., Qian, X.Y., Lu, H.Y.: Cross-sample entropy of foreign exchange time series. Physica A 389, 4785–4792 (2010)
23. Zhang, T., Yang, Z., Coote, J.H.: Cross-sample entropy statistic as a measure of complexity and regularity of renal sympathetic nerve activity in the rat. Exp. Physiol. 92, 659–669 (2007)
24. Ramdani, S., Seigle, B., Lagarde, J., Bouchara, F., Bernard, P.L.: On the use of sample entropy to analyze human postural sway data. Med. Eng. Phys. 31, 1023–1031 (2009)
25. van Schie, H.T., Waterschoot, B.M., Bekkering, H.: Understanding action beyond imitation: Reversed compatibility effects of action observation in imitation and joint action. J. Exp. Psychol. Human. 34, 1493–1500 (2008)

Part III
Brain Activity Measurement

Combined Linear Regression and Quadratic Classification Approach for an EEG-Based Prediction of Driver Performance

Gregory Apker, Brent Lance, Scott Kerick, and Kaleb McDowell

Human Research and Engineering Directorate
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD 21005, USA
{gregory.apker.ctr,brent.j.lance.civ,
scott.e.kerick.civ}@mail.mil
Kgm8@cornell.edu

Abstract. Electroencephalography (EEG) has been used to reliably and non-invasively detect fatigue in drivers. In fact, linear relationships between EEG power-spectral estimates and indices of driver performance have been found during simplified driving tasks. Here we sought to predict driver performance using linear regression in a more complex paradigm. Driver performance varied widely between participants, often varying greatly within a single driving session. We found that a non-selective linear regression model did not generalize well between periods of stable and erratic driving, yielding large errors. However, prediction errors were significantly reduced by training a linear regression model on stable driving for each participant. To provide a confidence estimate for the stable driving model, a quadratic discriminate classifier was trained to detect the transition from stable to erratic driving from the EEG power-spectra. Combined, the regression model and classifier yielded significantly lower prediction errors and provided improved discrimination of poor driving.

Keywords: EEG, Regression, Driving, Fatigue, Power Spectral Density.

1 Introduction

Fatigue and drowsiness are among the primary contributors to vehicular accidents, being estimated to have contributed to between 40-90% of all accidents [1-2]. In fact, a 2005 poll conducted by the National Sleep Foundation found that about 60% of adult drivers admitted to getting behind the wheel in a drowsy or fatigued state [3]. As a result, the prevention of these accidents has become a major focus of driver safety research.

To date, many systems have been designed to detect driver fatigue. Typically, these systems have relied on vehicle mounted sensors which correlate certain behaviors, such as vehicle dynamics, driver posture, or eye-blinking characteristics [4-6]. However, recent research has argued that monitoring the neural correlates of fatigue using electroencephalography (EEG) may provide a more reliable estimate of driver

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 231–240, 2013.

fatigue [7-8]. Further, a number of these studies have found significant correlations between neural signals and fatigue (see Lal and Craig, 2001, for a review [9]). Intriguingly, the results of these studies have varied almost as widely as their respective tasks, suggesting many differing assessments of the influence of fatigue on neural signals [10], leading to the conclusion that the specific influence of fatigue is task dependent [11-12].

Nonetheless, the observation of measureable changes in brain activity with fatigue has led to the development of several methods for classifying driver fatigue spanning a wide variety of classification approaches to predict the onset of fatigue [13-16], as well as discriminate between multiple levels of fatigue within a given driver [17-18]. While these works have been entirely fatigued-based, in a series of recent works, Lin and colleagues avoided the fatigue construct entirely and described a linear relationship between indices of driver performance and power-spectral estimates of EEG data [19-21]. In fact, they have shown that this simple relationship can be used to directly predict driver behavior based solely on neural activity recorded during a driving task with minimal processing of the EEG data [19].

However, the driving simulation used in their task was highly simplified. It has been shown that increases in task complexity can have a significant effect on the onset and characteristics of driver fatigue, and may be partly responsible for the diverse findings of the neural correlates of fatigue [11-12, 23]. As a result, it remains unclear how well a simple linear regression approach to driver performance prediction would translate to more complex driving tasks.

To begin to address this question, we evaluated predictions of driver performance from two linear regression models similar to that described in Lin et al. (2005a) in a more realistic driving scenario requiring participants to not only control vehicle heading but also control the speed of the vehicle and abide by posted speed limit signs. One of these models was trained on the full set of driving data during the training period while the other model only considered those points which reflect stable driving for that participant. Both models were then evaluated for the same testing data. We found significantly better performance of a linear model trained only on reasonably stable driving versus a linear model trained on the full range of behavior. Ultimately, we determined that the application of this type of performance prediction model benefits when coupled with an additional measure to diagnose changes in the relationship between power spectral estimates of EEG and driving behavior, thereby providing a confidence measure of the model prediction and insight into the driver's state.

2 Methods

2.1 Experimental Design

Participants. Eleven participants (aged from 20 to 40 years) participated in a virtual reality-based highway driving experiment. Each participant was briefed on the experimental equipment and procedures and signed an informed consent form. The voluntary, fully informed consent of the persons used in this research was obtained as required by Title 32, Part 219 of the Code of Federal Regulations and Army

Regulations 70-25. The investigator has adhered to the regulations for the protection of human participants as prescribed in AR 70-25.

Driving Simulation. Participants completed two separate driving sessions: the first, an acclimation session, lasted 15 minutes, the second experimental session consisted of 45 minutes of continuous driving. Before each session, participants provided an estimate of their fatigue level via the Karolinska Sleepiness Scale (KSS) [23]. Additionally, participants were asked to verbally report their fatigue score on this scale every 15 minutes during the second experimental session without interruption of driving.

Participants drove down a straight, infinitely long highway and were instructed to keep their vehicle as close to the center of the right-hand lane as possible. Throughout the session, after participants had maintained the vehicle within the appropriate lane for 8-10 seconds, a lateral perturbation was applied to the vehicle, causing it to begin to veer off course. The strength of the perturbation increased until the participant made a corrective steering adjustment (defined as a steering wheel deflection of 1 degree in the opposite direction of the perturbation) at which point the perturbation ceased allowing the participant to return the vehicle to center of the driving lane. The perturbation would ramp down automatically after approximately 3 seconds if no correction was made, however the participant was still required to correct the vehicle's heading and position. If the participant did not perform a corrective steering adjustment, the vehicle would continue to veer out of the lane and off the road until the vehicle was 21.9 meters outside of the lane, at which point the participant would be alerted to regain control of the vehicle via an auditory cue.

In addition to maintaining control of the vehicle's direction, participants also maintained appropriate speed for the vehicle during the testing session via accelerator and brake pedals. Participants were instructed to obey posted speed limit signs which appeared on the right-hand side of the road during the driving session. The speed limit was 45 mph for the majority of the session; however at three different points during the 45minute driving session the posted speed limit was reduced to 25 mph.

Data Collection and Analysis. Vehicle, EEG, and eye-tracker data were collected simultaneously throughout the experiment.

Vehicle Status and Performance Metrics. Vehicle status (position and dynamics) was monitored throughout each session, sampled at 90 Hz for participants 1-7 and at 100 Hz. for participants 8-11. To estimate driving performance, the vehicle's lateral deviation was calculated for entire session as the difference between the vehicle's lateral position and the center of the driving lane. To account for the tendencies of some participants to consistently position the vehicle to the right or left of the center of the lane, the median of their offset was subtracted to minimize this bias. Lane Deviation (LD) was then calculated as the absolute value of the lateral deviation throughout the driving session. LD values over the entire session were smoothed using a 90 second moving average filter with 2 second increments [19].

Electroencephalography. EEG signals were collected using a 64-channel Biosemi Active Two EEG system (Amsterdam, Netherlands), sampled at 2048 Hz and downsampled to 256Hz off-line. Electrode impedance was kept at or below 5 MΩ. Using

the embedded timing pulses and event signals, the EEG time series was synchronized with the vehicle status and driving performance data. Following this, the data was bandpass filtered to remove signals greater than 50 Hz and less than 5 Hz. The power spectral density estimates (PSD) for each channel were calculated using a 750 point Hanning window with 250 point overlap. Each channel and frequency power estimate of the 1-40 Hz bands was then smoothed with the same 90 second moving average filter used to smooth the lane deviation data, reducing variance and preserving the temporal alignment of the PSD and LD data streams. As was described in Lin et al. (2005a), correlation between PSD estimates and LD were often strongest for channels Cz and Pz leading to their selection for regression analysis. The same general trend was observed in the present study and thus the same two locations were used for performance prediction.

Eye tracking. Eye position was monitored but was not used for the analysis.

2.2 Experimental Design

Cross-Validation Preparation. The aligned EEG and vehicle data from the experimental session were split into three 15 minute blocks to train and test each prediction approach. Three-fold cross-validation was conducted such that two blocks were used to train the prediction algorithm and the remaining block was used to assess prediction performance. To eliminate overlapping data between training and testing sets, 90 seconds of the training data that abutted the testing data was removed prior to each cross-validation iteration.

Full-Data Regression. Following channel selection, principle component analysis (PCA) was then performed on the combined PSD estimates of both channels of the training session. Using these eigenvectors, both training and testing PSD estimates were projected into the component space and only the scores from the top 50 components (based on their eigenvalues) were preserved. The projected PSD data of the training set were used to calculate the coefficients of a 51 parameter (50 component vectors + offset) linear regression model of lane deviation. These coefficients were then applied to projected PSD testing data to generate a prediction of LD over this period. The predicted and measured LD values were compared for each epoch to characterize the predictive accuracy of the algorithm. This was repeated three times-- once for each cross-validation block.

Stable-Driving Regression. Driving performance varied widely not only between participants, but also within a single driving session for several of them. This high degree of variability resulted in dramatically different regression coefficients for a single participant's driving behavior depending upon which period of the driving session is used to train the model. To generate a regression model that yields more stable performance across an experimental session, we attempted to fit a linear regression model to a narrower subset of the driving performance data that reflected the more consistent driving epochs. To accomplish this, we defined a behavioral threshold for stable regression based on each participants individual driving habits:

$$"Stable\text{-}Driving"\ Threshold = \widetilde{LD}_i + 0.5\sigma_{LD,i} \quad (1)$$

where $\widetilde{LD}_i$ and $\sigma_{LD,i}$ are respectively the median and standard deviation of lane deviation during the training period for participant i. A linear regression model was subsequently calculated using only the LD data and PSD estimates from the indices in which LD values were below this threshold. In essence, LD values below this threshold represents a regime of stable driving performance; henceforth we refer to this sub-threshold performance as "stable driving".

Performance Classification and Confidence Estimate. Given that the "stable-driving" regression model was trained on a subset of the data when performance was generally consistent, the reliability of the model is somewhat limited during the periods of less stable driving where the shift in the behavior may be accompanied by a shift in the natural relationship between PSD estimates and driving performance. In these cases, the predictions of the stable-driving model may be suspect. It would be useful to be able to predict when the participant's performance may be deviating from stable driving based on the patterns observed in the PSD estimates, thereby providing an estimate of the confidence in the predictions of the stable-driving model. To accomplish this, a Quadratic Discriminate Classifier (QDC) was developed by assigning sub-threshold epochs to one class and supra-threshold epochs as another class. The PSD data fed into the classifier was treated identically to that used for regression with the one exception that the PSD estimates were smoothed with only a 4 second sliding window with 2 second steps. This was done to preserve a higher degree of sensitivity of the classifier to more rapid changes in PSD. Based on the PSD data of the testing period, if the QDC predicted stable driving conditions, we considered the predictions of the stable-driving model to be valid. However, if a transition to supra-threshold driving (class 2) was predicted by the QDC with 95% confidence, we considered the stable driving model's estimates to be invalid. In this way, the QDC serves as binary estimate of our confidence in the stable driving model.

Statistical Analysis. To compare predictive performance between models, regression coefficients, mean squared error (MSE) of the prediction, and identification of supra-threshold LD values within a given participants and block were compared directly using a paired Wilcox test unless otherwise stated. Significance threshold was set to a p-value of 0.05.

3 Results

3.1 Driving Performance

Driving performance varied greatly between participants as some participants maintained a high level of control of their LD whereas others exhibited periods of large LD or highly variable driving performance. For instance, 3 of the 11 participants' average LD did not exceed 0.5 meters, whereas 3 other participants produced averaged LD in excess of 2 full lanes outside of the correct lane.

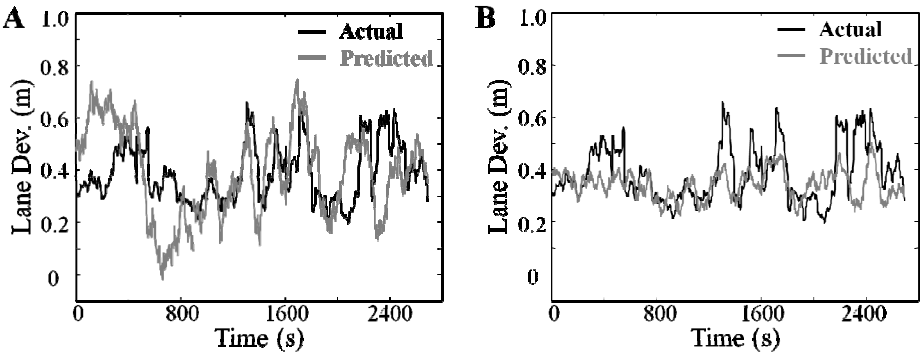


Fig. 1. Predicted Lane deviation for Full data and stable-driving data. Actual (black line) and predicted driver performance (grey lines) are drawn for a single driver for the full-data model **(A)** and the stable-driving model**(B)**.

3.2 Prediction of Driving Performance

Figure 1A illustrates the predicted LD for a single participant generated by the linear algorithm trained on the full set of driving data for each cross-validation block. For this participant, predicted performance varied widely across the driving session: In some portions of the session the predictions were generally accurate, while for other large portions, the prediction errors were very large. This trend was observed across all participants regardless of driving performance. The intra-session variability suggests that the linear relationship between the projected PSD data and driving performance was not consistent throughout the course of a single experiment.

The stable-driving model produced predictions of LD which were generally more accurate and less erratic across the entire session for each participant. Figure 1B illustrates the predicted LD from the stable-driving model for the same participant seen in Fig. 1A. In the figure, while the predictions of the model tended to under-estimate the full extent of driving error for large measured LD values, the predicted LD values appear to match the general trends and overall average LD levels. Importantly, this model did not produce large, inaccurate swings in predicted LD common in the full-data method. While predictions were generally more consistent across the experimental session using this approach, the correlation coefficients of model prediction did not significantly differ between approaches across the population.

The prediction errors of the full-data model increased as the variance increased in the training data. In fact, prediction errors of this model were significantly and positively correlated with the variability of the LD in the training data ($R=0.34$, $p< 0.05$), suggesting that model performance worsens when it is trained on a wider distribution of driving performance, perhaps because this additional variability cannot be explained by a single linear regression model. Interestingly, the prediction errors of the stable-driving model and the variability of training data were not found to be significantly correlated. Thus, in contrast to the full-data approach, the stable-driving model performance will not necessarily suffer for drivers who are more variable during the training period (or in general).

Comparing the prediction errors from each cross-validation block between models directly, the errors of the stable-driving approach were also significantly reduced relative to the original approach. The scatter plot shown in Fig. 2A indicates a significant bias for smaller prediction errors in the modified/stabilized algorithm. To quantify this, the relative improvement index (REI) was calculated as the ratio of the MSE of the full-data approach to that of the modified approach for each participant and block. An REI value greater than 1 indicates that the MSE of the full-data approach was greater than that for the stable-driving approach. As shown in Fig. 2B, a mean REI of 2.57 ± 2.83 and a median REI of 1.36 was observed and determined to be significantly greater than 1 (Wilcox signed rank test, $p<0.05$), indicating significantly smaller prediction errors in the stable-driving approach.

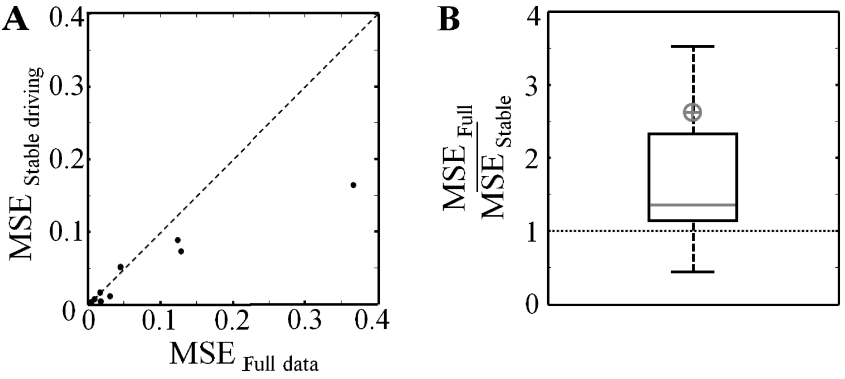


Fig. 2. Prediction error of Original vs. Modified algorithm. **A.** Scatter plot comparing the average MSE of the prediction for both linear algorithms for all participants. **B.** Relative Error Index as calculated from the MSE of each model, with mean (grey circle) and median (grey line) REI values.

3.3 Classification and Prediction of Driving Performance

In instances of large or variable LD values, the stable-driving model produced large prediction errors, indicating the linear relationship deteriorated during these periods. In these cases, the predicted LD from this model should no longer be trusted. A QDC was trained to detect changes in PSD which occurs beyond the stable-driving threshold to provide a confidence estimate for the stable-driving algorithm. Across all participants, the classifier was correctly identified the behavioral regime 88.3 ± 13% of the time. These class estimates serve primarily as an estimate of whether to trust the predictions of the stable-driving model. In the cases where the classifier predicted stable driving conditions from the PSD estimates, the average MSE of the stable-driving model predictions was 0.06 ± 0.14m across all participants. However, when the classifier predicted non-standard driving, the predicted LD of the stable driving model was always below the stable-driving threshold and had an average MSE of 0.17±0.3 m across the population, indicating the reduced confidence in the prediction was justified.

While the stable-driving model yielded significantly smaller predictions errors, the model often under-estimated large increases in LD and missed significantly more epochs of supra-threshold driving than did the linear model ($p<0.05$). This suggests the stable-driving model alone may not reliably predict when the participant begins to drive poorly. As previously described, the QDC output also serves as a predictor of supra-threshold driving. Combining the outputs of the classifier and stable-driving models may result in a more accurate estimate of sub- and supra-threshold driving epochs than the full-data model.

With respect to identifying periods of supra-threshold LD, the full-data model yielded predictive accuracy ranging from 56% to 100% across participants, with an average accuracy of 83.3 ± 15% for the population. The combined stable-driving and QDC system preformed better, yielding a range between 71% and 100%, with a significantly greater average accuracy of 89.8 ±11% ($p<0.05$). In addition, the number of false positives across the population, i.e. predictions of large LDs, was significantly greater in the full-data model compared to the combined approach ($p<0.05$), with no difference in the number of false negatives between approaches. Interestingly, the combined QDC and stable-driving predictions did not out-perform the QDC predictions of supra-threshold driving alone (88.6% average accuracy). This suggests that while the linear regression can provide a higher resolution estimate of the driving performance, the classifier was necessary to more reliably predict periods of the supra-threshold driving in this scenario, even when a more accurate linear model is used.

4 Discussion

In this study, we found that a linear algorithm indiscriminately trained on participants' driving data yielded larger predictions errors than one which was trained on a subset of driving data representative of stable-driving behavior. The stable-driving regression model was generally accurate during these periods of the testing data; however, performance deteriorated during periods of less stable driving. In this experiment, large or variable LDs were associated with a lack of vigilance on the part of the driver. Thus, the inability of the stable driving model to reliably predict far beyond the stable driving regime may be evidence of a shift in the natural relationship between PSD estimates and driving performance. This may in part explain why the model trained on the full set of data was less accurate, particularly in those cases where driving performance varied greatly.

Several researchers have recently applied non-linear algorithms to classify fatigue onset from EEG data with high degrees of accuracy [16-17], and in some cases discriminate between multiple levels of fatigue [18-19]; while others have shown broader network-based shifts in neural activity associated with fatigue driving performance [21]. Thus, it is possible that the onset of fatigue is accompanied by a more complex shift in the patterns of brain activity than can be characterized by a single linear algorithm. Another explanation for this is that additional processes or events not related to fatigue and drowsiness could have affected the relationship between PSD estimates and driver performance. Fatigue is only one of many physiological constructs which

can affect driving behavior and alter neural activity. A system designed to predict driver performance in the real-world must be equipped to manage or anticipate these factors to allay their affects.

Given this, we hypothesized that the predictions of a linear model may be complemented by a secondary means to detect when a shift in the relationship may occur. Here, a quadratic discriminate classifier was trained to detect a change in the patterns of PSD data indicative of a transition in behavior (i.e. from stable- to errant-driving) in and out of a regime where a single linear model could not extrapolate to. Using this output, we were able to identify epochs where the stable-driving model produced vastly larger errors and thus indicating that the QDC provided a useful confidence estimate for the stable-driving model. In addition, we used the classifier output as an additional behavioral metric and combined those predictions with the stable-driving to accurately predict periods of poor driving at a significantly higher rate than the linear model as well as produce significantly fewer false positives. As a result, we conclude that while a linear model trained on a limited regime of stable driving behavior yields improved predictions of driving behavior, this approach is greatly benefitted by a complementary method to specifically identify non-stable driving.

A potential future application for a regression/classification system as described here is to use the classifier to toggle between regression models based on the predicted state of the driver. That is, if a different relationship between PSD and LD is found during poor driving performance (or N other behavioral regime), the information provided by the classifier could also serve to switch between regression models trained specifically for the classified driver states. Further, the accuracy and reliability of such an EEG-based system may be enhanced by leveraging other sources of information regarding the driver's state such as eye-tracking, posture etc. This multimodal approach has been shown to be effective for detecting fatigue onset in drivers [15, 24] and would be of great benefit to an automated system deciding to trust the output of one or multiple predictions of driving performance.

References

1. Treat, J.R., Tumbas, N.S., McDonald, S.T., Shinar, D., Hume, R.D., Mayer, R.E., Stansifer, R.L., Castellan, N.J.: Tri-level study of the causes of traffic accidents. Report No. DOT-HS-034-3-535-77, TAC (1977)
2. Fletcher, K., McCulloch, S., Baulk, D., Dawson, D.: Countermeasures to driver fatigue: a review of public awareness campaigns and legal approaches. Aust. N.Z. J. Public Health 29, 471–476 (2005)
3. National Sleep Foundation. Sleep in America Poll, `http://www.sleepfoundation.org/article/sleep-america-polls/2005-adult-sleep-habits-and-styles`
4. Smith, P., Shah, M., da Vitoria Lobo, N.: Monitoring head/eye motion for driver alertness with one camera. In: Proc.15th International Conference on Pattern Recognition (ICPR 2000), Barcelona, Spain, vol. 4, pp. 636–642 (September 2000)
5. Perez, C.A., Palma, A., Holzmann, C.A., Pena, C.: Face and eye tracking algorithm based on digital image processing. In: Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC 2001), Tucson, Ariz, USA, vol. 2, pp. 1178–1183 (October 2001)

6. Popieul, J.C., Simon, P., Loslever, P.: "Using driver's head movements evolution as a drowsiness indicator. In: Proc. IEEE International Intelligent Vehicles Symposium (IV 2003), Columbus, Ohio, USA, pp. 616–621 (June 2003)
7. Okogbaa, O.G., Shell, R.L., Filipusic, D.: On the investigation of the nerophysiological correlates of knowledge worker fatigue using the EEG signal. Applied Ergonomics 25, 355–365 (1994)
8. Lal, S.K.L., Craig, A.: Driver Fatigue: Electroencephelography and psychological assessment. Psycholphyiology 29(3), 313–321 (2002)
9. Lal, S.K.L., Craig, A.: A critical review of the pyshcophysiology of driver fatigue. Biological Psychology 55, 173–194 (2001)
10. Craig, A., Tran, Y., Witjesurya, N., Nguyen, H.: Regional brain wave activity changes associated with fatigue. Psychophysiology 49, 574–582 (2012)
11. Desmond, P.A., Matthews, G.: Implications of task-induced fatigue effects for in-vehicle countermeasures to driver fatigue. Accid. Anal. Prev. 29(4), 515–523 (1997)
12. Desmond, P.A., Matthews, G.: Task-induced Fatigue Effects and Simulated Driving. Quart. Journal of Experimental Psychology 55(2), 659–686 (2002)
13. Peiris, M.T.R., Davidson, P.R., Bones, P.J., Jones, R.D.: Detection of lapses in responsiveness from the EEG. Journal of Neural Engineering 8 (2011)
14. Stikic, M., Johnson, R.R., Levendowski, D.J., Popovic, D.P., Olmstead, R.E., Berka, C.: EEG-derived estimators of present and future cognitive function. Frontiers in Human Neuroscience 5 (2011)
15. Sandberg, D., Akerstedt, T., Anund, A., Kecklund, G., Wahde, M.: Detecting Driver Sleepiness Using Optimized Non-Linear Combinations of Sleepiness Indicators. IEEE Trans. on Intelligent Transportation Systems 12(1), 97–108 (2011)
16. Zhoa, C., Zheng, C., Zhao, M., Tu, Y., Liu, J.: Multivariate autoregressive models and kernel learning algorithms for classifying driving mental fatigue based on electroencephalographic. Expert Systems with Applications 38, 1859–1865 (2011)
17. Shen, K.Q., Ong, C.J., Li, X.P., Wilder-Smith, E.P.V.: A feature selection method multilevel mental fatigue classification. IEEE Trans. Biomed. Eng. 54(7), 1231–1237 (2007)
18. Shen, K.Q., Li, X.P., Ong, C.J., Shao, S., Wilder-Smith, E.P.V.: EEG-based mental Fatigue measurement using multi-class support vector machines with confidence estimate. Clinical Neurophysiology 119, 1524–1533 (2008)
19. Lin, C.T., Wu, R.C., Jung, T.P., Liang, S.F., Huang, T.Y.: Estimating Driving Performance Based on EEG Spectrum Analysis. EURASIP Journal on Applied Signal Processing 19, 3165–3174 (2005a)
20. Lin, C.T., Wu, R.C., Liang, S.F., Huang, T.Y., Chao, W.H., Chen, Y.J., Jung, T.P.: EEG-based drowsiness estimation for safety driving using independent component analysis. IEEE Trans. Circ. Syst. 52, 2726–2738 (2005b)
21. Chuang, S.W., Ko, L.W., Lin, Y.P., Huang, R.S., Jung, T.P., Lin, C.T.: Co-modulatory spectral changes of independent brain processes are correlated with task performance. Neyroimage 62, 1467–1477 (2012)
22. Pattyn, N., Neyt, X., Henderickx, D., Soetens, E.: Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? Physiological Behavior 93(1-2), 369–378 (2008)
23. Akerstedt, T., Gillberg, M.: Subjective and objective sleepiness in the active individual. International Journal of Neuroscience 52, 29–37 (1990)
24. Yang, G., Lin, Y., Bhattacharya, P.: A driver fatigue recognition model based on information fusion and dynamic Bayesian network. Information Sciences 108, 1942–1954 (1942)

Differential Prefrontal Response during Natural and Synthetic Speech Perception: An fNIR Based Neuroergonomics Study

Hasan Ayaz[1,2], Paul Crawford[3], Adrian Curtin[1,2], Mashaal Syed[1,2], Banu Onaral[1,2], Willem M. Beltman[3], and Patricia A. Shewokis[1,2,4]

[1] School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, PA, 19104, USA
[2] Cognitive Neuroengineering and Quantitative Experimental Research (CONQUER) Collaborative, Drexel University, Philadelphia, PA, 19104, USA
[3] Intel Labs, Intel Corporation, Santa Clara, CA, 95054, USA
[4] Nutrition Sciences Department, College of Nursing and Health Professions, Drexel University, Philadelphia PA 19102, USA
hasan.ayaz@drexel.edu

Abstract. Synthetic speech has a growing role in human computer interaction and automated systems with the emergence of ubiquitous computing such as smart phones, car multimedia control and navigation systems. Cognitive processing costs associated with comprehension of synthetic speech relative to comprehension of natural speech have been demonstrated with behavioral (reaction time, accuracy, etc.) and self-reported (ratings, etc.) measures. In this neuroergonomics study, we have used optical brain imaging (fNIR: functional near infrared spectroscopy) to capture the brain activation of participants while they were listening to speech with varied quality, as well as natural speech. Results indicated a differential hemodynamic response with speech quality. As fNIR systems are safe, portable and record brain activation in real world settings, fNIR is a practical and minimally intrusive assessment tool for user experience researchers and can provide an objective metric for the design and development of next generation synthetic speech systems.

Keywords: Optical Brain Imaging, functional near infrared spectroscopy, fNIR, synthetic speech, perception, auditory processing.

1 Introduction

Speech perception is essential to human language and can be defined as the ability of a listener to identify and appropriately utilize the phonetic categories from audio stimuli input. It is known that separate prefrontal regions are specialized for the controlled processing of semantic information [1]. This study investigates cortical activation during different qualities of synthetic speech perception and processing as measured by optical brain imaging in human computer interaction settings.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 241–249, 2013.

Past research indicates that the human brain is wired and optimized for processing human voices (natural speech) and indicates that there are additional cognitive processing costs associated with processing and comprehension of synthetic speech [2-4]. Behavioral studies compare favorably for natural speech in comprehension although the difference might be slight in some cases. However, there currently is no tool available to user experience researchers to elicit objective measures of cognitive workload in ecologically valid environments.

Recent neuroimaging studies demonstrated brain activation changes with respect to quality (natural versus synthetic) of auditory stimuli with functional magnetic resonance imaging (fMRI) [4] and positron emission tomography (PET) [1] based studies. Both studies found significant differences in prefrontal cortex activity as a function of the quality of the auditory signal. Hence, these findings suggest that additional and complementary brain networks in the prefrontal cortex (cognitive processing) helps with covering modulation in auditory input quality to keep high performance (comprehension). Although a listener may not be even conscious to such processing, it yields longer reaction times, increased mental effort and eventually fatigue.

Validation and measurement of audio quality biomarkers and measurement using wearable optical brain imaging may help in the advancement of voice synthesizers and eventually assist in the production of synthetic speech that will sound more natural to and be easier to comprehend by human listeners.

The specific aim of this pilot study is to identify neural correlates of auditory processing and its relationship to stimuli quality as measured by functional near infrared (fNIR) spectroscopy which is a safe, non-invasive, affordable and portable neuroimaging technology that can be used to monitor hemodynamic changes that occur in the brain, i.e., blood oxygenation and blood volume, during select cognitive tasks such as mental workload [5-7], task difficulty/problem solving [8-10], performance[11-13] and learning[13-15] assessment tasks. Moreover, fNIR data can be collected in quiet settings unlike functional magnetic resonance imaging (fMRI) that exposes subjects to noise and confines them to restricted spaces and a supine position during the data acquisition process. These qualities pose fNIR as an ideal candidate for monitoring cognitive activity-related hemodynamic changes not only in laboratory settings but also under ecologically valid conditions – real world environments, consistent with the neuroergonomic [16] approach. A recent review of fNIR literature by Dieler et al. [17] summarizes the results of speech processing assessment in language-related disorders in the fields of neurology (i.e. aphasia and epilepsy) and psychiatry (i.e. disruptions of speech production in mood disorders, schizophrenia, dementia and anxiety disorders, as well as dyslexia and vigilance).

For the experimental paradigm, fNIR measures were integrated into a sentence listening task. The protocol involved listening to a sentence three consecutive times and rating the sentence each time for intelligibility, naturalness and overall quality in a balanced order. There were five different sentences (i.e., topics related to calendar information, email, navigation, sms and weather) and each were generated with three different quality levels and also recorded as natural speech being the highest quality. A 16-channel continuous wave (CW) fNIR system designed by the Optical Imaging Team at Drexel University (see [5]) was used to monitor the prefrontal cortex during task performance.

2 Methods

2.1 Participants

Four right-handed participants (assigned using the Edinburgh Handedness Inventory[18]) between the ages 22 to 25 volunteered for this study with average LQ of 72.64±17.13. Participants denied having hearing impairment, neurological or psychiatric history. All participants were medication-free, with normal or corrected-to-normal vision. Participants gave written informed consent for the study, which was approved by the Institutional Review Board at Drexel University, and were paid for their participation.

2.2 Experiment Protocol

The speech quality task and the synthetic speech recordings used in the study were originally developed at Intel Labs and were implemented as a Matlab application based on ITU Recommendation P.835 [19]. Before the task, a hearing test was performed for each participant and followed by a practice session with two trials before the task started. Each trial of the task started with listening to an audio recording (sentence) of about 5 seconds length. There were 5 different sentences and 4 levels of audio quality: natural (N) + 3 levels of synthetic (S1, S2 and S3). The synthetic speech synthesizers have different system requirements in terms of memory footprints. Synthesizer S1 required 250 MB, S2 required 1 MB and S3 required 50 MB. Participants were asked to rate intelligibility, naturalness and overall quality of the sound after listening to the each audio piece and submit selection from a scale of 1-bad, 2-poor, 3 fair, 4-good and 5-excellent. Stimulus delivery to the subject utilized a calibrated playback system with an auditory amplifier (Head Acoustics HPS IV), and high fidelity headphones (Sennheiser HD 600).

2.3 fNIR Data Acquisition

The continuous wave fNIR system (fNIR Devices LLC; www.fnirdevices.com) used in this study is connected to a flexible sensor pad that contains 4 light sources with built in peak wavelengths at 730 nm and 850 nm and 10 detectors designed to sample cortical areas underlying the forehead. With a fixed source-detector separation of 2.5 cm, this configuration generates a total of 16 measurement locations (optodes) [5, 20]. For data acquisition and visualization, COBI Studio software [21] (Drexel University) was used. The sampling rate of the system was 2Hz. During the task, a serial cable between the fNIR data acquisition computer and stimulus presentation computer was used to transfer time synchronization signals (markers) that indicate the start of sessions and onset of audio stimuli.

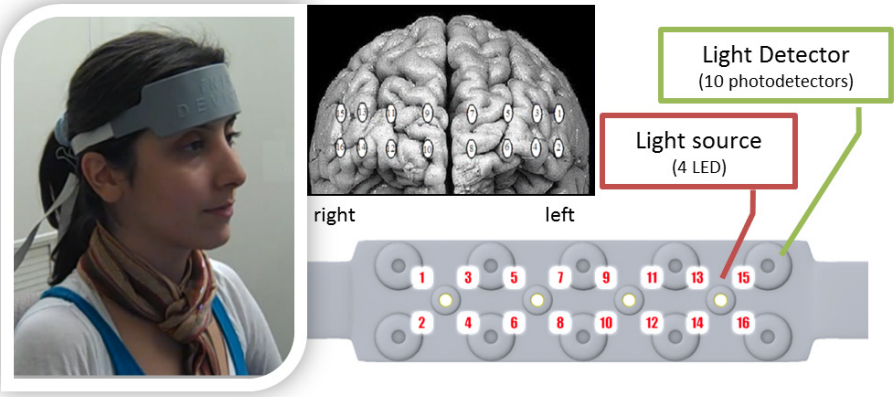


Fig. 1. Functional Near Infrared Spectroscopy sensor (head band) and optode locations visualized on anterior view brain surface image [5]

2.4 Data Analysis

For each participant, raw fNIR data was low-pass filtered with a finite impulse response, linear phase filter with order of 20 and cut-off frequency of 0.1Hz to attenuate the high frequency noise[5]. Motion artifact contaminated sessions and saturated channels (if any), in which light intensity at the detector was higher than the analog-to-digital converter limit were excluded[22]. Using time synchronization markers, fNIR data segments for rest periods (5 seconds before onset of audio) and task periods (audio file length plus 5 seconds) were extracted. Blood oxygenation changes within dorsolateral prefrontal cortex for all optodes were calculated using the Modified Beer Lambert Law (MBLL) for task periods with respect to rest periods at beginning of each task[5]. Average oxygenation change for each session was used as the dependent measure. For statistical analysis, one way repeated measures ANOVAs with 4 (Voice: Natural + 3 types of Synthetic) levels on both self-reported ratings and oxygenation changes were calculated for each optode. To account for violations of sphericity, Huynh Feldt corrections were used with Tukey Kramer post hoc tests to determine the locus of significant main effects. The significance criterion was 0.05. For multiple comparison correction, False Discovery Rate (FDR) approach was used [23]. This FDR based procedure has been reported to provide better balance between specificity and power than other available methods for multi-channel near-infrared spectroscopy functional neuroimaging data [24].

3 Results

3.1 Self-reported Measures

Subjective ratings recording during the first presentation of each sentence were submitted to a repeated measures one-way (speech quality) ANOVA. The self-reported ratings increase with the available amount of memory for respective recorded speech

synthesis solution. There was a significant main effect for speech audio quality ($F_{3,64}$=66.3, p<0.05) and is depicted in Fig. 2. Tukey post hoc tests indicated that each group was different from other ($q_{0.05/3,\ 9}$ =4.415, p<0.05) and showed that the natural speech had the highest ratings.

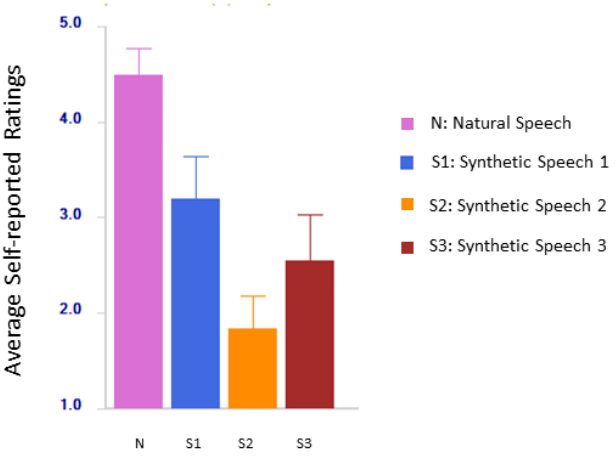


Fig. 2. Average self-reported measures for each speech quality level. Error bars are standard error of the mean (SEM)

3.2 fNIR Measures

Oxygenation values for all trials were submitted to a repeated measures one-way (speech quality) ANOVA for each optode separately, after FDR corrections only

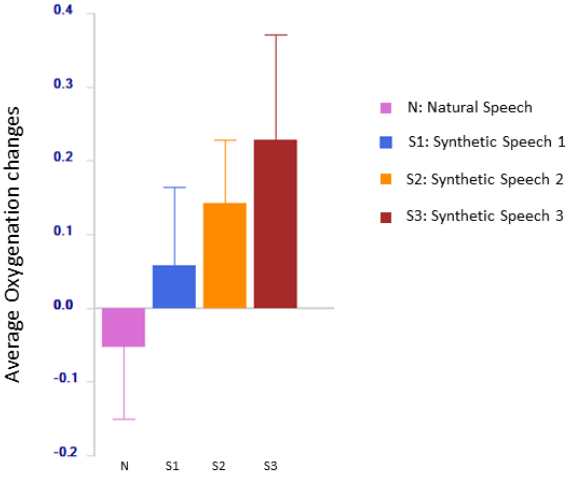


Fig. 3. Average task related oxygenation changes for all four speech quality levels. Error bars are standard error of the mean (SEM).

optode 13 (which approximately taps middle frontal gyrus) provided a significant response, depicted in Fig.3. There was a significant main effect for speech audio quality ($F_{3,44}$=10.27, p<0.05) and is depicted in Fig. 3. Tukey post hoc tests indicated that showed that the natural speech had significantly lower oxygenation compared to S2 and S3 ($q_{0.05/3,\ 9}$ =4.896, p<0.05).

3.3 Efficiency Analysis

Efficiency analysis provides a multidimensional view by connecting outcome and effort [25]. For this study, we estimated outcome with self-reported ratings since they indicated how much participants liked the speech inputs. For effort, we have used oxygenation changes as an indicator of cortical processing performed during that input as an objective assessment of cognitive effort. Normalization of both self-reported and oxygenation measures were performed by calculating z-scores with each subject separately.

In this efficiency graph, the fourth quadrant represents low efficiency, where minimum outcome is achieved with maximum effort. The second quadrant represents high efficiency where maximum outcome is achieved with minimal effort. The diagonal y=x is the neutral axis, where efficiency (E) is zero and effort and performance are equal. The Euclidian distance from the y=x axis (where E=0) indicates the efficiency for each condition. Efficiency graph for four speech quality levels (N, S1, S2 and S3) using all subjects' data is provided in Figure 4 below. To determine the relationship between normalized self-reported rating of speech audio quality and normalized oxygenation a zero-order correlation coefficient was calculated (r= -0.417, p <0.001). Also, calculated efficiencies for each four levels are listed within the graph.

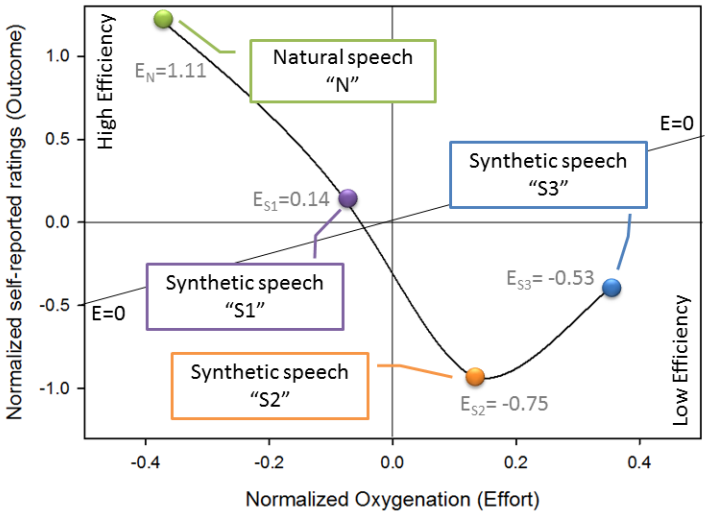


Fig. 4. Efficiency graph with normalized self-reported ratings (Outcome) vs. normalized oxygenation (Effort) graph for all four speech audio quality. Second quadrant is high efficiency and fourth quadrant is low efficiency area. y=x is where efficiency is zero.

4 Discussion

The purpose of this study was to test if cortical hemodynamic responses as measured by wearable optical brain imaging can detect a difference in prefrontal cortex with natural and synthetic speech. Our results indicate that oxygenation changes at optode 13, middle frontal gyrus had significant differences across varying levels of audio quality. In previous fMRI studies, a core group of regions beyond the auditory cortices, including the middle frontal gyrus has been shown to be preferentially activated for familiar speech categories and for novel non-speech audio [26]. Our results are also supporting the results from speech quality studies with fMRI and PET [1, 4].

Comparison of oxygenation changes with self-reported measures indicated a negative ($r = -0.417$) association between quality of speech and oxygenation: higher normalized oxygenation was observed for lower quality rated speech. Approximately, 17.4% of the variance in normalized self-reported speech quality output ratings can be explained by normalized oxygenation.

The efficiency analysis also provides insight into the relationship between the audio speech quality and neural activation representing cognitive effort. For example, oxygenation of natural speech was minimal, whereas the self-reported rating for it was highest and this resulted in highest efficiency rating as depicted in Figure 4. Furthermore, after applying the within subject normalization, the efficiency analysis indicated a transition from high efficiency (natural speech) to low efficiency (S2&S3).

Efficiency values (E) of each speech (N, S1, S2 and S3) that is the distance from y=x axis, followed the audio quality level; N had the highest and synthetic speeches followed the memory footprint used for the synthetic speech generation. The self-reported ratings and oxygenation values followed the same pattern, except for S2 and S3 oxygenation levels. Although the difference between oxygenation of S2 and S3 was not different, S2 had slightly lower oxygenation. One interpretation of this finding is that participants gave up processing the S2 at least in some trials perhaps by knowing that S2 had the lowest audio quality rather than the highest audio quality. However, additional experimentation is required to substantiate this speculation.

This study tested the effects of synthetic and natural speech in anterior prefrontal cortex and provides important albeit preliminary information about fNIR measures of the anterior prefrontal cortex hemodynamic response and its relationship to mental workload and speech perception. Level of audio quality does appear to influence the hemodynamic response in the dorsolateral/ventrolateral prefrontal cortices, at least for some complex sentences and with synthetic audio speech. Since fNIR technology allows the development of mobile, non-intrusive and miniaturized devices, it has the potential to be deployed in future human factors research environments to provide objective, task related brain-based measures of speech quality and may help in the design and development of complex human machine systems.

Acknowledgement. This study is made possible in part by a research award from the Intel Corporation and National Science Foundation (NSF) grant IIS:1065471. The content of the information herein does not necessarily reflect the position or the policy of the sponsors and no official endorsement should be inferred.

References

1. Sharp, D.J.: Monitoring and the Controlled Processing of Meaning: Distinct Prefrontal Systems. Cerebral Cortex 14, 1–10 (2004)
2. Hardee, J.B., Mayhorn, C.B.: Reexamining Synthetic Speech: Intelligibility and the Effects of Age, Task, and Speech Type on Recall. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 51, 1143–1147 (2007)
3. Paris, C.R., Thomas, M.H., Gilson, R.D., Kincaid, J.P.: Linguistic Cues and Memory for Synthetic and Natural Speech. Human Factors: The Journal of the Human Factors and Ergonomics Society 42, 421–431 (2000)
4. Benson, R.R., Whalen, D.H., Richardson, M., Swainson, B., Clark, V.P., Lai, S., Liberman, A.M.: Parametrically dissociating speech and nonspeech perception in the brain using fMRI. Brain Lang. 78, 364–396 (2001)
5. Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B.: Optical brain monitoring for operator training and mental workload assessment. Neuroimage 59, 36–47 (2012)
6. Girouard, A., Solovey, E.T., Jacob, R.J.K.: Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. International Journal of Autonomous and Adaptive Communications Systems 6, 26–44 (2013)
7. James, D.R.C., Orihuela-Espina, F., Leff, D.R., Sodergren, M.H., Athanasiou, T., Darzi, A.W., Yang, G.Z.: The ergonomics of natural orifice translumenal endoscopic surgery (NOTES) navigation in terms of performance, stress, and cognitive behavior. Surgery 149, 525–533 (2011)
8. Ayaz, H., Shewokis, P.A., İzzetoğlu, M., Çakır, M.P., Onaral, B.: Tangram solved? Prefrontal cortex activation analysis during geometric problem solving. In: 34th Annual International IEEE EMBS Conference, pp. 4724–4727. IEEE (2012)
9. Çiftçi, K., Sankur, B., Kahya, Y.P., Akın, A.: Functional Clusters in the Prefrontal Cortex during Mental Arithmetic. In: 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, pp. 1–4 (2008)
10. Hampshire, A., Thompson, R., Duncan, J., Owen, A.M.: Lateral prefrontal cortex subregions make dissociable contributions during fluid reasoning. Cerebral Cortex 21, 1–10 (2011)
11. Ayaz, H., Bunce, S., Shewokis, P., Izzetoglu, K., Willems, B., Onaral, B.: Using Brain Activity to Predict Task Performance and Operator Efficiency. In: Zhang, H., Hussain, A., Liu, D., Wang, Z. (eds.) BICS 2012. LNCS, vol. 7366, pp. 147–155. Springer, Heidelberg (2012)
12. Power, S.D., Kushki, A., Chau, T.: Towards a system-paced near-infrared spectroscopy brain-computer interface: differentiating prefrontal activity due to mental arithmetic and mental singing from the no-control state. Journal of Neural Engineering 8, 066004 (2011)
13. Ayaz, H., Cakir, M.P., Izzetoglu, K., Curtin, A., Shewokis, P.A., Bunce, S.C., Onaral, B.: Monitoring expertise development during simulated UAV piloting tasks using optical brain imaging. In: Aerospace Conference, 2012 IEEE, pp. 1–11 (2012)
14. Shewokis, P.A., Ayaz, H., Izzetoglu, M., Bunce, S., Gentili, R.J., Sela, I., Izzetoglu, K., Onaral, B.: Brain in the Loop: Assessing Learning Using fNIR in Cognitive and Motor Tasks. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) FAC 2011. LNCS, vol. 6780, pp. 240–249. Springer, Heidelberg (2011)
15. Pfurtscheller, G., Bauernfeind, G., Wriessnegger, S.C., Neuper, C.: Focal frontal (de)oxyhemoglobin responses during simple arithmetic. Int. J. Psychophysiol. 76, 186–192 (2010)

16. Parasuraman, R.: Neuroergonomics Brain, Cognition, and Performance at Work. Current Directions in Psychological Science 20, 181–186 (2011)
17. Dieler, A.C., Tupak, S.V., Fallgatter, A.J.: Functional near-infrared spectroscopy for the assessment of speech related tasks. Brain Lang. 121, 90–109 (2012)
18. Oldfield, R.C.: The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113 (1971)
19. ITU-T: Recommendation P.835 Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, vol. (11/2003). International Telecommunication Union, Geneva (2003)
20. Ayaz, H., Izzetoglu, M., Platek, S.M., Bunce, S., Izzetoglu, K., Pourrezaei, K., Onaral, B.: Registering fNIR data to brain surface image using MRI templates. In; Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 2671–2674 (2006)
21. Ayaz, H., Shewokis, P.A., Curtin, A., Izzetoglu, M., Izzetoglu, K., Onaral, B.: Using MazeSuite and Functional Near Infrared Spectroscopy to Study Learning in Spatial Navigation. J. Vis. Exp., e3443 (2011)
22. Ayaz, H., Izzetoglu, M., Shewokis, P.A., Onaral, B.: Sliding-window Motion Artifact Rejection for Functional Near-Infrared Spectroscopy. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 6567–6570 (2010)
23. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300 (1995)
24. Singh, A.K., Dan, I.: Exploring the false discovery rate in multichannel NIRS. Neuroimage 33, 542–549 (2006)
25. Paas, F.G.W.C., Van Merriënboer, J.J.G.: The efficiency of instructional conditions: An approach to combine mental effort and performance measures. Human Factors: The Journal of the Human Factors and Ergonomics Society 35, 737–743 (1993)
26. Husain, F.T., Fromm, S.J., Pursley, R.H., Hosey, L.A., Braun, A.R., Horwitz, B.: Neural bases of categorization of simple speech and nonspeech sounds. Human Brain Mapping 27, 636–651 (2006)

Functional Near-Infrared Spectroscopy in Addiction Treatment: Preliminary Evidence as a Biomarker of Treatment Response

Scott C. Bunce[1,2,3], Jonathan Harris[1], Kurtulus Izzetoglu[2,3], Hasan Ayaz[2,3], Meltem Izzetoglu[2,3], Kambiz Pourrezaei[2,3], and Banu Onaral[2,3]

[1] Penn State College of Medicine, Hershey, PA 17003 USA
[2] Drexel University School of Biomedical Engineering, Sciences, and Health Systems, Drexel University, Philadelphia, PA, 19104, USA
[3] Cognitive Neuroengineering and Quantitative Experimental Research Collaborative, Drexel University, Philadelphia, PA, 19104, USA
{sbunce,rmeyer,eobixler}@hmc.psu.edu, jdh@psu.edu,
{hasan.ayaz,meltem,ki25,kambiz,banu.onaral}@drexel.edu

Abstract. There is growing evidence that there are functional changes in the brains of individuals with substance use disorders. Numerous studies utilizing functional magnetic resonance imaging (fMRI) have shown that drug cues elicit increased regional blood flow in reward-related brain areas among addicted participants that is not found among normal controls. This finding has prompted leading investigators to suggest fMRI might be useful as a diagnostic or prognostic biomarker of addiction severity. However, fMRI is too costly for routine use in most treatment facilities. Functional near-infrared spectroscopy (fNIRs) offers an alternative neuroimaging modality that is safe, affordable, and patient-friendly. This manuscript reviews evidence that fNIRs can be used to differentiate prefrontal cortical responses of current alcohol dependent participants from alcohol dependent patients in treatment for 90-180 days. Differential responses to both alcohol and natural reward cues in both groups suggests fNIRs might serve as a clinic-friendly neuroimaging technology to inform clinical practice.

Keywords: Addiction, alcoholism, neuroimaging, functional near infrared spectroscopy, fNIRs, functional magnetic resonance imaging, fMRI, biomarker.

1 Introduction

A variety of techniques have been used to elucidate the pathophysiology of addiction, which includes abnormalities in brain structure, function, connectivity, and receptor pharmacology [1-3]. Recent neuroimaging studies have provided increasing evidence that there are indeed functional changes in the brains of individuals with substance use disorders (e.g., [2-4]). These functional changes have significant deleterious effects on people's behavior, and leave them at risk for continued substance abuse and its consequences.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 250–258, 2013.

Functional magnetic resonance imaging (fMRI) has been an important tool in the effort to understand the neurocircuitry underlying various aspects of addiction [1-3]. Numerous fMRI studies have now demonstrated that drug addicted individuals show increased regional blood flow in reward-related brain areas in response to drug cues that do not occur among normal controls. However, these studies have yet to be translated into clinically useful information that can be used to directly inform diagnosis or treatment. The consistency of fMRI-based studies on the functional differences between addicted participants and healthy controls, however, has led several prominent investigators to suggest that fMRI could be used as a biomarker of addiction severity [1, 4, 5] or treatment outcome [5]. Although fMRI is the current gold standard for non-invasive neuroimaging, and holds promise as a biomarker of addiction severity, the size, cost, and infrastructure required to operate an MRI system makes it untenable for use in a large majority of substance abuse treatment clinics.

Functional near-infrared spectroscopy (fNIRs), on the other hand, offers an affordable neuroimaging technology that could be readily implemented in a many clinical settings. fNIRs is a noninvasive optical imaging technique that can be used to monitor changes in the concentration of oxygenated hemoglobin (oxy-Hb) and deoxygenated hemoglobin (deoxy-Hb) during functional brain studies [6-10]. Analogous to fMRI, fNIRs provides information on local changes in blood oxygenation concentrations during neural activity, largely from the capillary beds. fNIRs can also be safely used for repeated measures on the same individual. In contrast to fMRI, however, fNIRs can be engineered to provide neuroimaging systems that are relatively inexpensive, portable, boast rapid application time (5-10 minutes), and near-zero run-time costs. fNIRs is also relatively robust to movement artifacts in comparison to fMRI, allowing more ecologically valid experimental paradigms. Participants can be sitting and working at a computer, standing, even walking on a treadmill while being monitored with fNIRs. Algorithms have been developed to remove motion artifacts should they occur during desktop as well as ambulatory use [11-13]. Having an affordable neuroimaging technology that can be implemented in a typical clinical office makes it feasible for routine clinical use at drug and alcohol treatment centers.

fNIRs does have two important limitations relative to fMRI. First, with greater depth of penetration, there is an exponential decrease in the amount of light that scatters back to the surface of the scalp. Given the magnitude of oxygenation changes associated with cognitive/emotional activity, this limits current fNIRs neuroimaging to the outer cortex (2-3 cm) of the brain [14]. Second, due to the scattering properties of light interacting with biological tissue, spatial localization is on the order of 1 $cm^{2,}$ versus the 1-2mm^2 of fMRI. Despite these limitations, if a given phenomenon of interest is located in accessible cortex, fNIRs provides the potential for safe, comfortable, affordable, and portable neuroimaging.

The utility of fNIRs in addiction medicine derives from two emerging themes in the addiction literature; first, that the dorsolateral/ventrolateral prefrontal cortex plays an important role in the individual's response to both drug cues and natural rewards, and second, that anhedonia, or failure to respond to natural rewards, plays a critical role in relapse among patients in treatment for addiction. It is well-established that the

reinforcing effects of drugs of abuse are mediated by the meso-corticolimbic dopaminergic system [15,16]. A large number of neuroimaging studies of cue reactivity have identified a distributed neural network that is activated by drug- and alcohol-related stimuli among participants with drug addiction [1, 17-21]. Until recently, theories of addiction focused primarily on reward processes mediated by mesolimbic dopaminergic circuits (e.g., [22]). However, recent studies suggest that dorsolateral pre-frontal cortex (DLPFC), orbitofrontal cortex (OFC) and anterior cingulate cortices, comprise a mesocortical dopamine circuit involved in behavioral control mechanisms as well as in the conscious experience of drug intoxication [1, 23]. Drug cues are known to be perceived as highly appetitive by non-treatment seeking opiate addicts [24, 25]. Drug users selectively attend to drug-related cues at the expense of other stimuli (e.g., [26]), and attention is largely supported by dorsolateral prefrontal areas [27]. Furthermore, in their 2004 review of the cued response literature, Wilson et al. [18] suggested that differential activation in areas of DLPFC and OFC in response to drug cues may be related to treatment status. Among neuroimaging studies that examined non-treatment seeking individuals, 8 of 10 found activation in DLPFC, whereas only 1 of 9 studies that examined treatment-seeking individuals found activation in either DLPFC or OFC.

In contrast, non-addicted individuals preferentially respond to natural reward cues rather than drug cues. Whereas drug users are inclined to perceive drug-related cues as positively valenced, non-users are not [28-30]. Neuroimaging studies have shown that the long-term use of drugs of abuse decreases dopamine (DA) striatal D2 receptors and DA release [17], resulting in diminished responses to natural rewards. Because the large and long-lasting increases in DA induced by drugs of abuse are still able to activate the compromised reward circuits, whereas natural reinforcers are not, the salience of drug cues over natural reinforcers is thought to fuel relapse [17,31]. These attentional and evaluative biases are posited to operate automatically, outside awareness, and to exert a controlling influence over drug-taking behavior [32]. As such, anhedonia, or the inability to experience natural rewards as reinforcing, is gaining as a central construct in our understanding of relapse.

Theoretically, the prefrontal cortex plays a critical role in the integration of motivational and cognitive information, and in mediating the neural basis for adaptive processing of incentive stimuli [18]. It is also involved in the assignment of emotional significance to a stimulus and producing an affective state in response [33]. These observations parallel developing models of DA that view it not only as a neurotransmitter of reward, but also as playing a role in signaling the salience of events (including aversive, rewarding, novel, and unexpected stimuli), in driving motivation, in predicting reward - or failure to receive it, and in facilitating memory consolidation of salient events [1, 17, 23]. In light of this research, recent theories have begun to emphasize the critical role of cortical function in drug abuse (e.g., [1]). Goldstein & Volkow [1, 23] have proposed a model that conceptualizes drug addiction as a syndrome of impaired response inhibition and salience attribution. In their model, the core of drug addiction is a loss of self-directed, volitional behaviors to automatic processes driven by the primary need for drug in lieu of other rewarding stimuli. Disruption of prefrontal top-down processes (mediated by dopaminergic

processes) releases behaviors that are typically tightly monitored and regulated. If human drug addiction, indeed, down-regulates the frontal cortex and its supervisory functions, the role of higher cognitive and self-monitoring processes in addiction are critical to our understanding of relapse prevention. Growing evidence for the role of prefrontal cortex in addiction, coupled with research suggesting that neuroimaging can be used to predict relapse [34-37], makes fNIRs a viable neuroimaging technology that could readily be used in the clinical office, or even in a bar setting, to provide an objective measure of diagnostic or prognostic utility.

2 Functional Near-Infrared Spectroscopy in Addiction Research

Bunce et al [38] utilized fNIRs to evaluate the hypothesis generated by Wilson et al. [18], i.e., that current alcohol-dependent participants with no motivation to stop drinking would show increased activation in DLPFC/OFC to alcohol cues relative to patients who had been in treatment for 90-180 days and social drinkers. They also evaluated the participants' responses to natural reward cues, predicting reduced response to reward cues among current drinkers relative to patients in treatment and social drinkers.

2.1 Prefrontal Responses to an Alcohol Cued Response Task

The methods for the study are presented in detail elsewhere [38]. In brief, participants in the study were 14 right-handed non-smokers recruited into three groups; 4 nontreatment-seeking adult alcoholics (1 female), 6 alcoholic patients currently in recovery (2 females), and 4 healthy social drinkers (2 females). Diagnoses were assigned using the Structured Clinical Interview for *DSM–IV* for Axis I (Ver. 2.0), and daily alcohol use for the 180 days prior to intake were gathered using the Form-90 A interview [39]. NTSA met DSM-IV criteria for Alcohol Dependence, expressed no interest in treatment, and had not sought treatment in the past year. RA met DSM-IV criteria for Alcohol Dependence in early full remission, lived in a non-restricted environment, and reported no alcohol use for 90-180 days. This pattern of sobriety was the behavioral operationalization of early commitment to sobriety, as they reported having remained sober past the critical early (90 day) phase of relapse [40-42], while having had the opportunity to drink. Social drinkers reported consuming fewer than 7 drinks per week. All participants registered a Blood Alcohol Content (BAC) of .000 (Alco-Sensor IV), prior to imaging, and scored 1 or less on the Clinical Institute Withdrawal Assessment for Alcohol-revised (CIWA-Ar; [43]).

Participants were asked to complete a cued response task. Visual stimuli were presented in a block design, with each block consisting of either: a) alcoholic beverages, b) nonalcoholic beverages, c) visual control pictures, d) a crosshair, or e) natural rewards (highly palatable food). The alcohol blocks were specific to a beverage type (wine, beer, or liquor), with two blocks per type. After each block, participants rated their craving and resistance to craving in real time on 100-point

visual analog scales. fNIRs sensors were located over bilateral dorsolateral and inferior frontal gyri [44].

The results showed that, as predicted, current drinkers had increased activation to alcohol cues over right middle/inferior frontal gyrus relative to participants in treatment as well as social drinkers (see Fig 1; F (2,11) = 7.62, p = .008; partial eta2 = .58). Patients in treatment showed marginally less activation to the alcohol cues than the social drinkers. The results were reversed in response to the natural reward cues.

Current drinkers had significantly less neural activation in response to the natural reward cues than either participants in recovery or the social drinkers, whereas social drinkers and patients in recovery did not differ (Fig. 2). This effect was also found in the right hemisphere, slightly more posterior, towards inferior frontal gyrus relative to the area activated by the alcohol cues (Fig. 3).

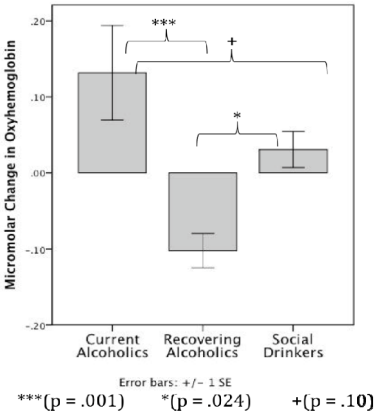


Fig. 1. Mean changes in OxyHb in response to to Alcohol stimuli

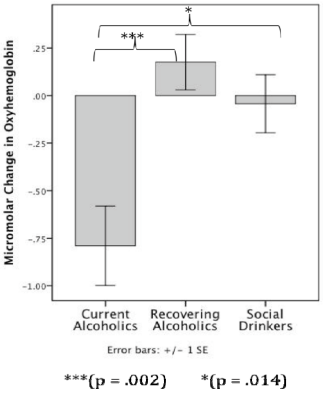


Fig. 2. Mean changes in OxyHb in response natural reward stimuli

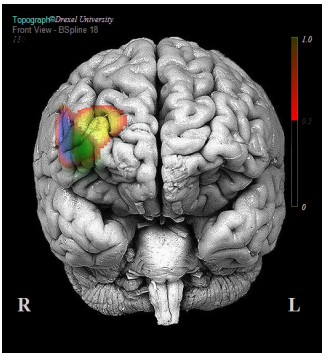


Fig. 3. Location of Neural Response to Alcohol Cues and Natural Rewards. Yellow = activation to alcohol – beverage cues; Blue = activation to natural rewards – beverage cues; Green = overlap in activation to both alcohol and natural reward cues (Optode 14).

Similar results using fNIRs imaging have been reported for patients who are in treatment for prescription opiate dependence. In a preliminary report, Bunce et al. [45] suggested that patients who had just completed detoxification for prescription opiates (n=7) showed increased activation to prescription pill cues in right lateral prefrontal cortex relative to patients who had been in extended supervised residential treatment for 60-90 days (n=7). The area of activation in response to the pill cues had substantial overlap with the results from the Bunce et al. study in alcoholics, although the effect size was not as large. The smaller effect size may have been due to the fact that the latter study looked at patients at two different stages of early treatment, rather than current users versus patients in extended sobriety. Both studies, however, require larger sample sizes and further elaboration.

3 Conclusions

The findings in these fNIRs studies are consistent with growing evidence from research employing fMRI [1, 2, 23] indicating that prefrontal cortices are involved in the cycle of addiction. Bunce et al. [38] found that current drinkers showed a heighted response to alcohol cues in lateral prefrontal cortex relative to patients in sustained recovery from alcohol dependence and normal controls. In contrast, natural reward cues elicited greater responses in right lateral prefrontal cortex from participants in recovery, and decreased responses from non-treatment seeking participants. Although the exact interpretation of these cortical responses still remains to be determined, increased activation among current users and patients in early recovery in this area of the cortex make it likely that it is related to attentional processes. Attentional biases towards drug-related stimuli, and a lack of attention to natural reinforcers, have been well documented in addiction [1,2]. These findings are consistent with Goldstein and Volkow's [1, 23] impaired response inhibition and salience attribution (iRISA) model of addiction, which argues that disrupted prefrontal cortical function leads to a syndrome in which addicted individuals attribute excessive salience to the drug and drug-related cues, coupled with decreased sensitivity to non-drug reinforcers, and a decreased capacity to inhibit disadvantageous or maladaptive behaviors.

The differential response to alcohol and natural reward cues in both current alcoholics and patients in extended sobriety is important for two reasons. First, although this was a cross-sectional study, rather than a longitudinal study, it suggests that the hedonic response to natural rewards, if compromised in addiction, may return with extended sobriety. Deficient response to non-drug related rewards is a known problem in treatment, and a critical factor in the addiction cycle [e.g., 1, 21], as drugs of abuse remain the primary source of gratification among patients in early recovery. An objective, brain-based measure of a patient's hedonic capacity would be helpful to improve treatment planning. Second, this finding answers an important potential criticism of the Bunce et al. study, i.e., that the current drinkers had imbibed alcohol much more recently than the patients in recovery. The differential cortical responses to alcohol and natural reward cues suggest that the results cannot be attributed to a general hypometabolism in cortical response among either group. Both current and recovering alcoholics had cortical responses to relevant stimuli, but to psychologically different stimuli.

There are other limitations to these studies. First, given the small sample sizes, these results must be interpreted with caution until larger studies can be completed. Second, the cortical area that was assessed in these studies was limited, which in turn limits the capacity to fully understand the implications of the data. More research is necessary, including studies that integrate the results of fNIRs and fMRI, to fully explicate the meaning and clinical utility of these preliminary results.

In conclusion, the research reviewed in this manuscript suggests that, like fMRI, fNIRs may have utility as a biomarker of addiction severity, or as a prognostic indicator of relapse vulnerability in addition treatment. fNIRs has the added potential to provide affordable and patient-friendly neuroimaging for routine clinical use in treatment facilities that do not have access to an fMRI magnet, and for research in ecologically valid environments such as a bar setting. If, indeed, fNIRs can provide an objective index of predilection to relapse, clinicians could use it to develop better treatments through the use of an objective biomarker, and provide better care through individualized medicine.

References

1. Goldstein, R.Z., Volkow, N.D.: Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. Nat. Rev. Neurosci. 12, 652–669 (2011)
2. Koob, G.F., Volkow, N.D.: Neurocircuitry of Addiction. Neuropsychopharmacol 35, 217–238 (2010)
3. Reske, M., Paulus, M.P.: The diagnostic and therapeutic potential of neuroimaging in addiction. In: Bryon, A., Stein, E.A. (eds.), pp. 319–343. John Wiley & Sons, Ltd., Chichester (2011)
4. O'Brien, C.: A potential biomarker for addiction. Neuropsychopharmacol 38, S17 (2012)
5. Volkow, N.D., Wang, G.J., Fowler, J., Tomasi, D.: Addiction circuitry in the human brain. Annu Rev. Pharmacol. Toxicol. 52, 321–336 (2012)
6. Boas, D.A., Gaudette, T., Strangman, G., Cheng, X., Marota, J.J.A., Mandeville, J.B.: The accuracy of near infrared spectroscopy and imaging during focal changes in cerebral hemodynamics. Neuroimage 13, 76–90 (2001)
7. Chance, B., Zhuang, Z., UnAh, C., Alter, C., Lipton, L.: Cognition-activated low-frequency modulation of light absorption in human brain. Proc. Natl. Acad. Sci. USA 90, 3770–3774 (1993)
8. Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K., Li, C., Murray, T., Ovetsky, Y., Pidikiti, D.: A novel method for fast imaging of brain function, non-invasively, with light. Opt. Express 2, 411–423 (1998)
9. Obrig, H., Villringer, A.: Near-infrared spectroscopy in functional activation studies. Can NIRS demonstrate cortical activation? Adv. Exp. Med. Biol. 413, 113–127 (1997)
10. Villringer, A., Chance, B.: Non-invasive optical spectroscopy and imaging of human brain function. Trends Neurosci. 20, 435–442 (1997)
11. Ayaz, M., Izzetoglu, P., Shewokis, A., Onaral, B.: Sliding-window Motion Artifact Rejection for Functional Near-Infrared Spectroscopy. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 6567–6570 (2010)
12. Izzetoglu, K.: Neural correlates of cognitive workload and anesthetic depth: fNIRs spectroscopy investigation in humans. (Doctoral Dissertation) Retreived from Drexel E-Repository and Archives (iDEA) (2008), http://hdl.handle.net/1860/2896

13. Izzetoglu, M., Chitrapu, P., Bunce, S., Onaral, B.: Motion artifact cancellation in NIR spectroscopy using discrete Kalman filtering. Biomed. Eng. Online 9, 16 (2010)
14. Chance, B., Leigh, J., Miyake, H., Smith, D., Nioka, S., Greenfeld, R., Finander, M., Kaufmann, K., Levy, W., Young, M.: Comparison of time-resolved and-unresolved measurements of deoxyhemoglobin in brain. Proc. Natl. Acad. Sci. USA 85, 4971 (1988)
15. Koob, G.F., Le Moal, M.: Drug abuse: Hedonic homeostatic dysregulation. Science 278, 52–58 (1997)
16. Bossert, J.M., Ghitza, U.E., Lu, L., Epstein, D.H., Shaham, Y.: Neurobiology of relapse to heroin and cocaine seeking: An update and clinical implications. Eur. J. Pharmacol. 526, 36–50 (2005)
17. Volkow, N.D., Fowler, J.S., Wang, G.J.: The addicted human brain viewed in the light of imaging studies: Brain circuits and treatment strategies. Neuropharmacology 47(suppl.1), 3–13 (2004)
18. Wilson, S.J., Sayette, M.A., Fiez, J.A.: Prefrontal responses to drug cues: A neurocognitive analysis. Nat. Neurosci. 7, 211–214 (2004)
19. Liu, C., Showalter, J., Grigson, P.S.: Ethanol-induced conditioned taste avoidance: Reward or aversion? Alcohol Clin. Exp. Res. 33, 522–530 (2009)
20. Yang, Z., Xie, J., Shao, Y.C., Xie, C.M., Fu, L.P., Li, D.J., Fan, M., Ma, L., Li, S.J.: Dynamic neural responses to cue-reactivity paradigms in heroin-dependent users: An fMRI study. Hum. Brain Mapp. 30, 766–775 (2009)
21. Zijlstra, F., Veltman, D.J., Booij, J., van den Brink, W., Franken, I.H.: Neurobiological substrates of cue-elicited craving and anhedonia in recently abstinent opioid-dependent males. Drug Alcohol. Depend 99, 183–192 (2009)
22. Blum, K., Braverman, E.R., Holder, J.M., Lubar, J.F., Monastra, V.J., Miller, D., Lubar, J.O., Chen, T.J., Comings, D.E.: Reward deficiency syndrome: A biogenetic model for the diagnosis and treatment of impulsive, addictive, and compulsive behaviors. J. Psychoactive Drugs 32(suppl. i-iv), 1–112 (2000)
23. Goldstein, R.Z., Volkow, N.D.: Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. Am. J. Psychiatry 159, 1642–1652 (2002)
24. Robinson, T.E., Berridge, K.C.: The neural basis of drug craving: An incentive-sensitization theory of addiction. Brain Res. Brain Res. Rev. 18, 247–291 (1993)
25. Robinson, T.E., Berridge, K.C.: Addiction. Annu. Rev. Psychol. 54, 25–53 (2003)
26. Townshend, J.M., Duka, T.: Attentional bias associated with alcohol cues: Differences between heavy and occasional social drinkers. Psychopharmacol 157, 67–74 (2001)
27. Cabeza, R., Nyberg, L.: Imaging cognition II: An empirical review of 275 PET and fMRI studies. J. Cogn. Neurosci. 12, 1–47 (2000)
28. Mucha, R.F., Geier, A., Pauli, P.: Modulation of craving by cues having differential overlap with pharmacological effect: Evidence for cue approach in smokers and social drinkers. Psychopharmacol 147, 306–313 (1999)
29. Mogg, K., Bradley, B.P., Field, M., De Hower, J.: Eye movements to smoking-related pictures in smokers: Relationship between attentional biases and implicit and explicit measures of stimulus valence. Addiction 98, 825–836 (2003)
30. Field, M., Mogg, K., Zetteler, J., Bradley, B.P.: Attentional biases for alcohol cues in heavy and light social drinkers: The roles of initial orienting and maintained attention. Psychopharmacol 176, 88–93 (2004)
31. Kenny, P.J., Polis, I., Koob, G.F., Markou, A.: Low dose cocaine self-administration transiently increases but high dose cocaine persistently decreases brain reward function in rats. Eur. J. Neurosci. 17, 191–195 (2003)

32. Childress, A.R., Ehrman, R.N., Wang, Z., Li, Y., Sciortino, N., Hakun, J., Jens, W., Suh, J., Listerud, J., Marquez, K., Franklin, T., Langleben, D., Detre, J., O'Brien, C.P.: Prelude to passion: limbic activation by "unseen" drug and sexual cues. PLoS One 3, 1506 (2008)
33. Phillips, M.L., Drevets, W.C., Rauch, S.L., Lane, R.: Neurobiology of emotion perception I: The neural basis of normal emotion perception. Biol. Psychiatry 54, 504–514 (2003)
34. Grüsser, S.M., Wrase, J., Klein, S., Hermann, D., Smolka, M.N., Ruf, M., Weber-Fahr, W., Flor, H., Mann, K., Braus, D.F.: Cue-induced activation of the striatum and medial prefrontal cortex is associated with subsequent relapse in abstinent alcoholics. Psychopharmacol 175, 296–302 (2004)
35. Heinz, A., Wrase, J., Kahnt, T., Beck, A., Bromand, Z., Grüsser, S.M., Kienast, T., Smolka, M.N., Flor, H., Mann, K.: Brain activation elicited by affectively positive stimuli is associated with a lower risk of relapse in detoxified alcoholic subjects. Alcoholism: Clin Exp. Res. 31, 1138–1147 (2007)
36. Janes, A.C., Pizzagalli, D.A., Richardt, S.: Brain reactivity to smoking cues prior to smoking cessation predicts ability to maintain tobacco abstinence. Biol Psychiatry 67, 722–729 (2010)
37. Paulus, M.P., Tapert, S.F., Schuckit, M.A.: Neural activation patterns of methamphetamine-dependent subjects during decision making predict relapse. Arch Gen Psychiatry 62, 761–768 (2005)
38. Bunce, S.C., Izzetoglu, K., Izzetoglu, M., Ayaz, H., Pourrezaei, K., Onaral, B.: Treatment Status Predicts Differential Prefrontal Cortical Responses to Alcohol and Natural Reinforcer Cues among Alcohol Dependent Individuals. In: Zhang, H., Hussain, A., Liu, D., Wang, Z. (eds.) BICS 2012. LNCS, vol. 7366, pp. 183–191. Springer, Heidelberg (2012)
39. Miller, W.R.: Form 90: A Structured Assessment Interview for Drinking and Related Behaviors (Test Manual). NIAAA Project MATCH Monograph Series, vol. 5. NIH Publication No. 96-4004. Natl. Inst. Alcohol Abuse Alcoholism, Bethesda (1996)
40. Miller, W.R., Hester, R.K.: Inpatient alcoholism treatment: Who benefits? Am. Psychol. 41, 794 (1986)
41. Emrick, C.D.: A review of psychologically oriented treatment of alcoholism: I. The use and interrelationships of outcome criteria and drinking behavior following treatment. Quar. J. Stud. Alcohol. 35, 523–549 (1974)
42. Hunt, W.A., Barnett, L.W., Branch, L.G.: Relapse rates in addiction programs. J. Clin. Psychol. 27, 455–456 (1971)
43. Sullivan, J.T., Sykora, K., Schneiderman, J., Naranjo, C.A., Sellers, E.M.: Assessment of Alcohol Withdrawal: the revised clinical institute withdrawal assessment for alcohol scale (CIWA-Ar). Brit. J. Addiction 84, 1353–1357 (1989)
44. Okamoto, M., Dan, H., Sakamoto, K., Takeo, K., Shimizu, K., Kohno, S., Oda, I., Isobe, S., Suzuki, T., Kohyama, K.: Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10-20 system oriented for transcranial functional brain mapping. Neuroimage 21, 99–111 (2004)
45. Bunce, S.C., Bixler, E.O., Harris, J., Meyer, R.E.: A clinical laboratory model of allostasis of the brain reward system diurnal cortisol and sleep in recently detoxified opioid dependent patients, normal control subjects & patients drug free for 60-90 days. Neuropsychopharmacol 38, S444–S445 (2012)

Towards Noise-Enhanced Augmented Cognition

Alexander J. Casson

Department of Electrical and Electronic Engineering,
Imperial College London, SW7 2AZ, UK
acasson@imperial.ac.uk

Abstract. Workload classification Augmented Cognition systems aim to detect when an operator is in a high or low workload state, and then to modify their work flow and operating environment based upon this knowledge. This paper reviews state-of-the-art electroencephalography (EEG) recorders for use in such systems and investigates the impact of EEG noise on an example system performance. It is found that adding up to 15 μV_{RMS} of artificially generated noise still leaves EEG signals that have correlations in-line with the correlations found between conventional wet EEG electrodes and new dry electrodes. The workload classification system is found to be robust in the presence of small amounts of noise, and there is initial evidence of small stochastic resonance effects whereby better performance can actually be obtained in the noisy case compared to the traditional noise-less case.

Keywords: EEG, Augmented Cognition, Workload classification, Noise-enhanced signal processing.

1 Introduction

Augmented Cognition is a recent research concept focusing on creating the next generation of Human-Computer Interaction devices. Closed-loop Brain Computer Interfaces (BCIs) are a classic example of such next generation systems. In these, a human operator uses a computer and interacts with changes on the screen; whilst simultaneously the computer monitors the human and changes its outputs based upon the results. For example, workload monitoring systems aim to detect when an operator is in a high or a low workload state, and use this knowledge to change the speed at which information is presented to the operator. As such the work flow and operating environment can be optimized in a real-time and time-varying manner.

Successful BCI Augmented Cognition intrinsically relies on the availability of portable and easy-to-use brain monitoring technologies. For this there are two practical modalities, functional near-infrared (fNIR) and electroencephalography (EEG). The EEG is the non-invasive recording of *brainwaves* performed non-invasively by placing electrodes on the scalp, and is by far the most commonly used modality. As a result, in recent years there has been a huge amount of research dedicated to improving the EEG unit and the overall recording experience. [1]–[6] represent a small selection of such papers.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 259–268, 2013.

Although both have seen considerable process in recent years the two principle focuses in EEG unit research are well known, and remain: power consumption and dry electrode design. In Section 2 this paper presents a brief review of state-of-art EEG technology for use in Augmented Cognition, highlighting the recent improvements on these two fronts. An in-depth analysis on the impact of recording noise on Augmented Cognition performance is then presented in Section 3. Excess noise in the EEG recording is related to the use of dry electrodes through the correlation coefficients obtained as *clean* EEG signals are corrupted by artificially generated noise. By injecting small amounts of artificial noise into the EEG collected from a workload monitoring task it is shown that the task performance is robust under noisy EEG recordings. Further, initial evidence of small stochastic resonance effects, where the system performance actually improves in noisy conditions, is found.

2 Portable EEG for Augmented Cognition

2.1 EEG Recorders

Table 1 summarises the features of state-of-the-art low channel count EEG systems that are potentially suitable for non-obtrusive EEG brain monitoring in Augmented Cognition applications. Low channel counts are sufficient for many applications, and for Augmented Cognition the need for recorders that are discrete, socially acceptable, and quick to set up, places a strong emphasis on the use of a low number of channels.

From Table 1 it can be seen that a number of high quality, highly miniaturised units are now available commercially. These can easily offer over 8 hours of recording time, likely sufficient for any individual protocol in an Augmented Cognition experiment. Nevertheless, one day of recording, allowing a complete sleep-wake cycle to be captured, should be the aim for future high-quality units. (In any case, even the best clinically attached wet electrodes begin to fall off after this time.) This 24 hour level of power consumption is starting to be met by research stage units.

However this still falls far short of *pick up and use* devices. Substantial improvements in system power consumptions will be required to realise units that can be trusted to be re-usable session after session. Although current batteries guarantee that a wanted protocol is feasible, it remains a common experience to have to worry about battery charge, or to have to adjust experiment timings after discovering that a unit was not adequately charged. Tackling this is essential for engendering user trust and reliability in Augmented Cognition systems. Onboard signal processing for providing the first level analysis of the EEG data is a promising approach for further power consumption reductions, but implementing complete and accurate algorithms within the limited power budget available remains a major challenge [1].

Looking further ahead, the EEG technology itself is evolving. For example, [4] reported the use of very small, flexible, textile based EEG units. These are applied directly to the scalp as a *tattoo* and, if forehead only channels are required,

Table 1. Approximate specifications of state-of-the-art low channel count EEG systems for use in Augmented Cognition. Many devices come in different models and configurations; only one potential configuration is reported here. Physical sizes are as given by the manufacturer and are not directly comparable: some are for the recorder unit alone while others are for the complete EEG system.

Device	Channels	Sampling frequency / Hz	Resolution / bits	Size / mm	Weight / g	Battery life / hours	Wireless?	Dry electrodes?	Status
Actiwave [7]	4	128	8	37× 27× 8.5	8.5	13	No	No	Commercial
Emotiv [8]	14	128	14	–	116	12	Yes	No	Commercial
B-Alert [9]	4	256	16	127× 57× 25	110	8	Yes	No	Commercial
NeuroSky [10]	1	512	12	225× 115× 165	90	8	Yes	Yes	Commercial
Sleep zeo [11], [12]	1	128	12	–	24	8 (1 night)	Yes	Yes	Commercial
Enobio [13]	8	500	24	225× 115× 165	65	8	Yes	Yes	Commercial
Quasar [14]	12	240	16	–	500	24	Yes	Yes	Commercial
IMEC [15], [16]	8	1000	12	35× 30× 5	100	22	Yes	Yes	Research
MINDO [17]	4	512	16	165× 145× 50	100	20	Yes	Yes	Research

eliminate much of the wiring involved in the EEG collection and are very inconspicuous. [18] presented a new approach for recording the EEG from the ear canal using a modified hearing aid. This is a very interesting development because the recording location is accessible, it intrinsically holds the electrodes in place, and hearing aids are already very socially acceptable. It also allows a single unit that can collect free-running EEG and auditory steady state responses, while simultaneously collecting a heartbeat record and providing classic hearing aid functionality. Both of these developments are at an early stage, but hold significant promise for future use in Augmented Cognition applications.

2.2 Electrode Technologies

Also apparent from Table 1 is the increasing availability of dry EEG electrodes which do not require a conductive gel to operate. Most of these electrodes are now based upon having *fingered* electrodes, rather than *discs*, for easier penetration through the hair (see for example [19]). It is clear that making a fundamentally gel free recording is no longer a major challenge. However, there are outstanding challenges in how to actually keep the electrodes in place without a cap or tight headband. Furthermore, electrode availability does not mean that these electrodes get comparable performance to conventional wet Ag/AgCl EEG recording electrodes.

In-depth measurements of dry electrode performance have been presented [6], [20], [21] but most studies only report a correlation coefficient between EEG recorded at nearby locations with wet and dry electrodes. Typical values reported are: >0.93 [3]; 0.89 [22]; 0.83 [23]; 0.81–0.98 [15]; 0.68–0.90 [16]; 0.39–0.85 [24]. For greater acceptance of dry electrodes the wider reporting of the second order electrode properties is essential. In particular: the half-cell potential, the long term stability and the contact noise. The latter is known to be a function of electrode contact area [25], which is decreasing with the move to fingered electrodes. To begin to evaluate the impact of this, the remainder of this paper investigates the effect of excess recording noise on a workload monitoring Augmented Cognition task.

3 Noise-Enhanced Augmented Cognition

3.1 Noise Correlation

Noise robustness is a clear requirement of Augmented Cognition systems that must operate in non-controlled environments. Excess recording noise from any source cannot be allowed to have a substantial detrimental effect on the system performance. To investigate this, Fig. 1 shows the correlation coefficient calculated between a raw recorded EEG trace and the same EEG trace after it has had artificial white Gaussian noise deliberately added to it. The additive noise generation procedure is detailed in [26]. In Fig. 1, the artificial noise is added to a complete 12.5 hour EEG recording (using the publicly available data from [27],

[28]). This long EEG record is then split into multiple shorter duration EEG sections, and the correlation in each section plotted against the duration of these shorter sections. This allows the maximum, minimum and median correlation coefficients over time to be found.

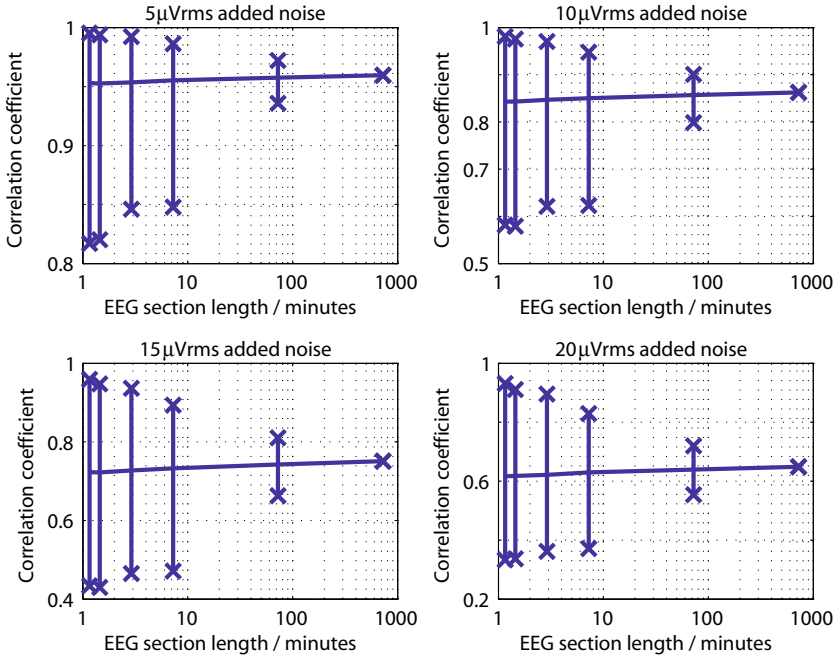


Fig. 1. Correlation coefficients between a raw EEG trace and a noise corrupted copy of the same EEG trace as the EEG section length used for calculation is changed. Vertical lines show the maximum, minimum and median correlation values found over a complete 12.5 hour EEG recording.

From Fig. 1 it is clearly seen that the underlying correlation present is not accurately estimated when very short sections of data are analysed. There is a consistent tendency for the median correlation to be underestimated at the cost of much larger variances. As a result, in some cases only testing the correlation in short EEG records will lead to a significant overestimation of the true correlation present. Importantly, even with up to 15 μV_{RMS} of artificial noise added to the raw EEG traces, correlations in-line with those reported for dry electrodes are found.

It is therefore essential to investigate the impact of this noise on Augmented Cognition system performance. Moreover, recent results have shown that some EEG applications are not only robust in the presence of more noise, but actually get better performance [26]. Such *stochastic resonance* has been observed in many physical systems [29] and could have a big impact on EEG in Augmented

Cognition. For example, is it necessary to design electrodes to have the minimum contact noise anyway?

3.2 Noise-Enhanced Processing

These effects are investigated here using an EEG workload classification system based upon the publicly available data from the 2011 Cognitive State Assessment Competition [30], [31]. In this, participants were asked to perform a workload engagement task [32], [33] which altered the difficultly and required attention level between high and low workload states. Nineteen channels of EEG data were recorded, and the experiment was run on each person multiple times on the same day, and on different days. The objective is to use only the EEG data to recognise the operator's state as either high or low workload.

Fig. 2 shows the performance of a new Artificial Neural Network based workload monitor on the data from two subjects. The used network is a simple feed-forward patternnet with 10 hidden neurons with features from standard FFT frequency bands and time domain features including line-length. These are calculated from all 19 EEG channels. The used Artificial Neural Network is trained using the first recording session from day 1. The test data is then taken as the two other recording sessions on day 1, and the three from day 2. Fig. 2 shows that the Artificial Neural Network performs well on day 1, the same day as the

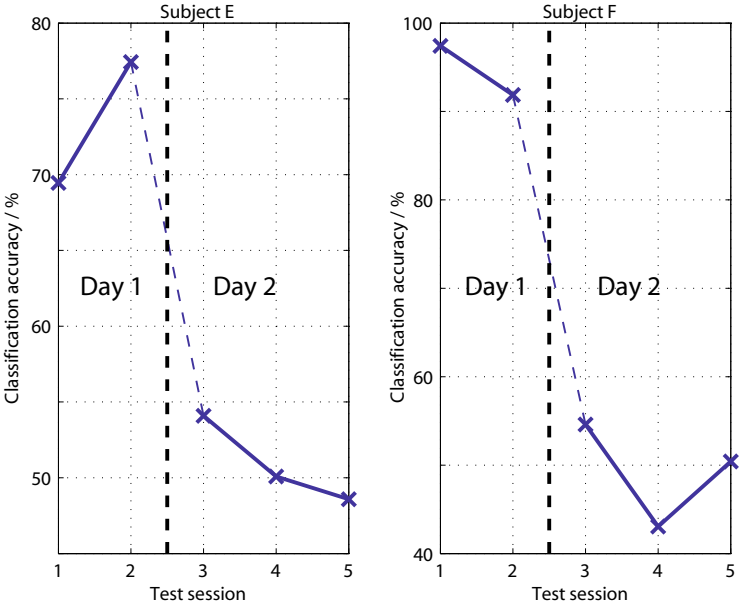


Fig. 2. Performance of an Artificial Neural Network workload monitor using data from two subjects recorded on two subsequent days. Data is taken from the 2011 Cognitive State Assessment Competition [30], [31].

training data is from. However, by day 2 (the next day) the network performance has degraded substantially and is no better than chance.

This result, using a different Artificial Neural Network, replicates the results reported in [30], [31] which demonstrated that the performance of some workload classification systems degraded significantly as the time gap between the training and testing sessions increased. Clearly such systems are not reliable and reusable. Re-training of the network is required each day and this comes with a high time cost. There are now open research questions over the causes of these performance decreases, and potential approaches for mitigating them.

The impact on this situation from adding artificially generated noise to the raw EEG traces is shown in Fig. 3. *Training with noise* is a common technique used to increase the accuracy of Artificial Neural Networks by adding small levels of noise to the training data before training the network [34]. The aim is to do this multiple times and make the available training data more variable and more representative of future unknown data. *Testing with noise* is a novel approach introduced here where independently generated noise is also added to the EEG data used for testing. This therefore simulates the use of a more noisy EEG recorder for obtaining the test data. It also simulates the potential use of low-power, low-accuracy circuit structures in the EEG unit in place of conventional higher-accuracy, higher-power structures. As such the noise results here are useful for creating even low power consumption EEG processing electronics.

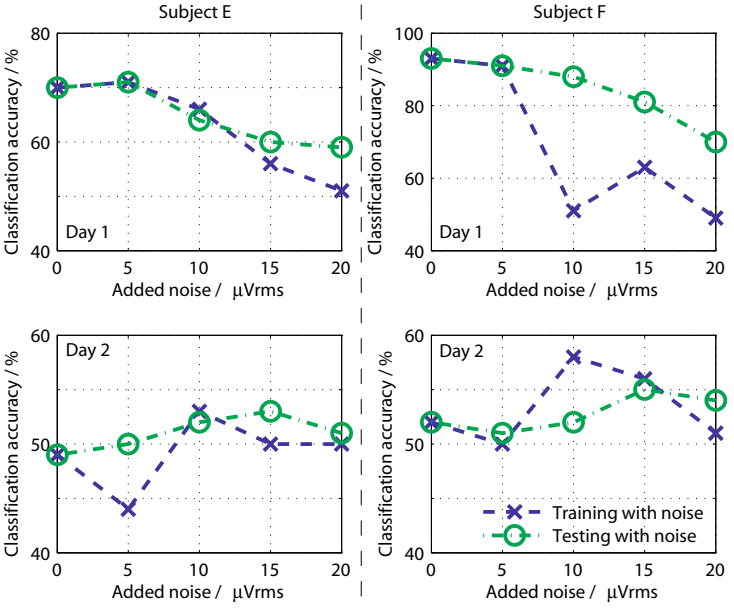


Fig. 3. Performance of an Artificial Neural Network workload monitor as artificial noise is deliberately added to the training and test data. Plotted results show the average performance the test sessions on each day (two on day 1 and three on day 2).

From Fig. 3, in both subjects the presence of excess noise in the EEG recording does not intrinsically stop the workload classification process. Robust performance is maintained when small amounts of noise are present. Moreover, several instances of performance improvements are present. Considering day 1, in Subject E a small performance resonance is present with better classification accuracies being obtained when 5 μV_{RMS} of noise is deliberately added to the EEG signals. In Subject F no resonance is seen, but there is no substantial decrease in performance. On day 2, better classification performance is obtained at many different noise levels compared to the no noise case. This effect is small, and the issue with performance degradation over time is not fixed: in neither of the cases considered here does the performance improve to a level substantially above chance classification. Nevertheless, this demonstration of stochastic resonance effects is an important new result for Augmented Cognition systems. If this effect can be isolated and improved upon, noise enhanced processing could be an important new tool for creating robust and reusable Augmented Cognition systems that can work autonomously over a number of days.

4 Conclusions

Stochastic resonance is an effect whereby noise embedded in a signal leads to better overall performance compared to a no noise case. This paper has demonstrated that EEG systems are now readily available with dry EEG electrodes for quick and easy set ups. These electrodes produce EEG signals with high correlations when compared to conventional wet electrodes, but similar correlations can be obtained when using EEG signals which have been artificially corrupted by up to 15 μV_{RMS} of noise. Using an EEG based Artificial Neural Network workload classification system as an example this paper has shown that the system performance is maintained under such noise levels. Indeed there is initial evidence of stochastic resonance effects, with consistently better performance being obtained on next day workload classifications tests as more noise is added to the EEG data. At present these stochastic resonance effects are very small, but suggestive, and future work investigate their full exploitation.

References

1. Casson, A.J., Yates, D.C., Smith, S.J., Duncan, J.S., Rodriguez-Villegas, E.: Wearable electroencephalography. IEEE Eng. Med. Biol. Mag. 29, 44–56 (2010)
2. Verma, N., Shoeb, A., Bohorquez, J., Dawson, J., Guttag, J., Chandrakasan, A.P.: A micro-power EEG acquisition SoC with integrated feature extraction processor for a chronic seizure detection system. IEEE J. Solid-State Circuits 45, 804–816 (2010)
3. Xu, J., Yazicioglu, R.F., Grundlehner, B., Harpe, P., Makinwa, K.A.A., Van Hoof, C.: A 160 μW 8-channel active electrode system for EEG monitoring. IEEE Trans. Biomed. Circuits Syst. 5, 555–567 (2011)

4. Kim, D.H., Lu, N., Ma, R., Kim, Y.S., Kim, R.H., Wang, S., Wu, J., Won, S.M., Tao, H., Islam, A., Yu, K.J., Kim, T.I., Chowdhury, R., Ying, M., Xu, L., Li, M., Chung, H.J., Keum, H., McCormick, M., Liu, P., Zhang, Y.W., Omenetto, F.G., Huang, Y., Coleman, T., Rogers, J.A.: Epidermal electronics. Science 333, 838–843 (2011)
5. Nikulin, V.V., Kegeles, J., Curio, G.: Miniaturized electroencephalographic scalp electrode for optimal wearing comfort. Clin. Neurophysiol. 121, 1007–1014 (2010)
6. Chi, Y., Jung, T.P., Cauwenberghs, G.: Dry-contact and noncontact biopotential electrodes: Methodological review. IEEE Rev. Biomed. Eng. 3, 106–119 (2010)
7. cam*n*tech Actiwave: Home page (2013), http://www.camntech.com/
8. Emotiv EEG systems: Home page (2013), http://www.emotiv.com/
9. Advanced Brain Monitoring B-Alert X4: Home page (2013), http://advancedbrainmonitoring.com/
10. NeuroSky MindWave: Home page (2013), http://www.neurosky.com/
11. Sleep Zeo: Home page (2013), http://www.myzeo.com/sleep/
12. Shambroom, J.R., Fabregas, S.E., Johnstone, J.: Validation of an automated wireless system to monitor sleep in healthy adults. J. Sleep Res. 21, 221–230 (2012)
13. Neuroelectrics Enobio: Home page (2013), http://neuroelectrics.com/
14. Quasar DSI 10/20: Home page (2013), http://www.quasarusa.com/
15. IMEC: Holst centre and panasonic present wireless low-power active-electrode EEG headset (2012), http://www.imec.be/
16. Patki, S., Grundlehner, B., Verwegen, A., Mitra, S., Xu, J., Matsumoto, A., Yazicioglu, R.F., Penders, J.: Wireless EEG system with real time impedance monitoring and active electrodes. In: IEEE BioCAS, Hsinchu (2012)
17. Mindo 4H Earphone: Home page (2013), http://www.mindo.com.tw/
18. Looney, D., Kidmose, P., Park, C., Ungstrup, M., Rank, M.L., Rosenkranz, K., Mandic, D.P.: The in-the-ear recording concept: User-centered and wearable brain monitoring. IEEE Pulse 3, 32–42 (2012)
19. g.tec g.sahara: Home page (2013), http://www.gtec.at/
20. Slater, J.D., Kalamangalam, G.P., Hope, O.: Quality assessment of electroencephalography obtained from a "dry electrode" system. J. Neurosci. Methods 208, 134–137 (2012)
21. Gandhi, N., Khe, C., Chung, D., Chi, Y.M., Cauwenberghs, G.: Properties of dry and non-contact electrodes for wearable physiological sensors. In: Int. Conf. BSN, Dallas (2011)
22. Matthews, R., McDonald, N.J., Hervieux, P., Turner, P.J., Steindorf, M.A.: A wearable physiological sensor suite for unobtrusive monitoring of physiological and cognitive state. In: IEEE EMBC, Lyon (2007)
23. Gargiulo, G., Bifulco, P., Calvo, R.A., Cesarelli, M., Jin, C., van Schaik, A.: A mobile EEG system with dry electrodes. In: IEEE BioCAS, Baltimore (2008)
24. Estepp, J.R., Christensen, J.C., Monnin, J.W., Davis, I.M., Wilson, G.F.: Validation of a dry electrode system for EEG. In: Proc. HFES, San Antonio (2009)
25. Huigen, E., Peper, A., Grimbergen, C.A.: Investigation into the origin of the noise of surface electrodes. Med. Biol. Eng. Comput. 40, 332–338 (2002)
26. Casson, A.J., Rodriguez-Villegas, E.: Utilising noise to improve an interictal spike detector. J. Neurosci. Methods 201, 262–268 (2011)
27. De Clercq, W., Vergult, A., Vanrumste, B., Van Paesschen, W., Van Huffel, S.: Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. IEEE Trans. Biomed. Eng. 53, 2583–2587 (2006)

28. Vergult, A., De Clercq, Q., Palmini, A., Vanrumste, B., Dupont, P., Van Huffel, S., Van Paesschen, W.: Improving the interpretation of ictal scalp EEG: BSS-CCA algorithm for muscle artifact removal. Epilepsia 45, 950–958 (2007)
29. Kay, S.: Can detectability be improved by adding noise? IEEE Signal Processing Lett. 7, 8–10 (2000)
30. Christensen, J.C., Estepp, J.R., Wilson, G.F., Russell, C.A.: The effects of day-to-day variability of physiological data on operator functional state classification. Neuroimage 59, 57–63 (2012)
31. Estepp, J.R., Klosterman, S.L., Christensen, J.C.: An assessment of non-stationarity in physiological cognitive state assessment using artificial neural networks. In: IEEE EMBC, Boston (2011)
32. Comstock, J.R., Arnegard, R.J.: The multi-attribute task battery for human operator workload and strategic behavior research. Technical report, TM-104174, National Aeronautics and Space Administration Langley Research Center (1992)
33. Miller Jr., W.D.: The U.S. air force-developed adaptation of the multi-attribute task battery for the assessment of human operator workload and strategic behavior. Technical report, AFRL-RH-WP-TR-2010-0133, U.S. Air Force Research Laboratory (2010)
34. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, Oxford (1995)

Soft, Embeddable, Dry EEG Sensors for Real World Applications

Gene Davis, Catherine McConnell, Djordje Popovic,
Chris Berka, and Stephanie Korszen

Advanced Brain Monitoring, Inc. 2237 Faraday Ave., Ste 100 Carlsbad, CA 92008
Gene@b-alert.com

Abstract. Over the last decade, numerous papers have presented the use of dry electrodes capable of acquiring electroencephalogram (EEG) signals through hair. A few of these dry electrode prototypes have even progressed from lab-based EEG acquisition to commercial sales. While the field has improved rapidly as of late, most dry electrodes share a number of shortcomings that limit their potential real world applications including: 1) multiple rigid prongs that require sustained pressure to penetrate hair and maintain solid scalp contact, creating higher levels of discomfort when compared to standard wet sensors; 2) cumbersome or chin-strap-type applications for maintaining electrode contact, creating barriers to end user acceptance; 3) rigid active electrodes to compensate for high input impedances that limit flexibility and placement of sensors; 4) inability to safely imbed sensors under protective headgear, restricting use in some fields where EEG metrics are most desired; and 5) expensive sensor manufacturing that drives costs high for use across subjects. Under a recent DARPA Phase 3 contract, Advanced Brain Monitoring has developed a novel semi-dry sensor that addresses the current dry electrode shortcomings, opening up the door for new real world applications without compromising subject safety or comfort. The semi-dry sensor prototype was tested during a live performance requirement at the end of Phase 3, and successfully acquired EEG across all subject hair types over a 3 day testing period. The results from the performance requirement and subsequent results for new advancements to the prototype are presented here.

Keywords: Electroencephalograms (EEG), dry-electrodes, wearable EEG, BCI, Real World Applications.

1 Introduction

While Electroencephalography (EEG) has been used for decades to record the electrical activity of the brain [1] and validated for use in a wide range of applications it has rarely left the controlled confines of the laboratory. Use of medical-grade EEG in real world settings has often been limited by its susceptibility to environmental noise, usability constraints, and availability of technical personnel. Over the last twenty years, however, technological advances have begun to address these issues, enabling medical grade, real world wearable systems.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 269–278, 2013.

The "Holy Grail" for EEG is a self-applied, wearable system that can reliably record medical grade EEG on users in the real world. In the quest for this ideal system, the notion of a “Dry EEG Sensor” has become a popular buzz word, and (in some cases) a de facto requirement. This trend can be explained in part by the unfavorable way in which most dry sensor publications portray "wet" sensor technologies [2-7], particularly as it pertains to Brain Computer Interface (BCI) platforms. The bulk of dry sensor publications use older EEG systems that have since been updated or superseded as their examples of "current" wet EEG platforms. This perpetuates the misconception that the only wet sensor systems available require substantial time to set up each sensor site, depend upon extensive skin preparation (below 5kΩ), are not wearable, are susceptible to electromagnetic interference due to leads from the head to the amplifier, and result in severe discomfort to the user [2-3]. Some combination of these qualities are often listed as shortcomings of existing wet sensors, and thereby benefits of implementing dry sensors. Dry sensor publications often further emphasize their advantage by overstating the amount of residue left behind by wet sensors. These publications rarely acknowledge the existing available medical grade wet sensors that are multi-site systems with short set up times, minimal or zero skin preparation, easily attainable impedances below 80kΩ, wearability for multiple days (during both wake and sleep), low electromagnetic interference in wireless mode and/or storage directly to the device, and high levels of user comfort [8-11]. One remaining, frequently cited, drawback specific to wet sensors is the residue left behind. As aforementioned, dry electrode publications commonly reference wet systems that use 10/20 paste and collodian, and require an experienced laboratory technician for application. In reality, some current wet systems have already eliminated any residues, and for many other applications the residue is minimal and unnoticeable. Moreover, depending on intended use, many wet sensor systems can be self-applied by the end user without any technical personnel required [12-13], and those applications requiring assistance can easily be completed by non-technical personnel.

What remains as a significant drawback for wet sensors when compared against dry sensors is the ability to record long acquisitions (i.e., over 8 hours) without requiring the addition of more gel or paste. This dry sensor benefit should, however, be considered alongside the negatives inherent in the current state of dry electrodes, to include: 1) multiple rigid prongs that require sustained pressure to penetrate hair and maintain solid scalp contact, creating higher levels of discomfort when compared to standard wet sensors; 2) cumbersome or chin-strap-type applications for maintaining electrode contact, decreasing the likelihood of end user acceptance; 3) rigid active electrodes to compensate for high input impedances that limit flexibility and placement of sensors; 4) inability to safely imbed sensors under protective headgear, restricting use in some fields where EEG metrics are most desired; and 5) expensive sensor manufacturing that drives costs high for use across subjects. When objectively evaluating existing dry sensors vs. current wet technologies, intended application should be taken into consideration. For short term recordings (i.e., less than 8 hours), evidence suggests that available wet sensors are the more effective option, while for

acquisitions over 8 hours in length dry sensors may prove beneficial. Another viable option for future wearable sensors is to extend the 8 hour recording time of existing wet sensors.

Under a DARPA Phase 3 contract, we were able to consider both possibilities as possible end solutions for long acquisitions. Solution 1 entailed the design of a dry hydrogel sensor that was soft, flexible, and embeddable, eliminating all of the main drawbacks associated with existing dry sensor technologies. The design process allowed the dry sensor to be interchangeable with our current wet (i.e., foam and synapse cream) sensors, while maintaining the same usability across head sizes and hair types. Some of the results from a 9 subject study are included, along with additional single subject studies on the most recent advances and modifications of the dry sensor.

Solution 2 involved improving the ease-of-use of the current wet systems to ultimately enable the end user to self-adjust the system, quickly and easily changing out the sensors as needed without any additional support. This development would permit long term recordings with wet sensors. Some early prototype solutions are highlighted in the Discussion section.

2 Methods and Materials

2.1 Methods

As part of the Phase 3 DARPA contract, the Advanced Brain Monitoring, Inc. (ABM) dry sensor prototype was integrated with a proprietary EEG system and tested as part of a live performance requirement. A total of 9 subjects (1 female; ages: 22- 39) participated in the study, with each of the 3 dry electrode teams providing 3 subjects. The subjects rotated through a 3 day testing sequence across all 3 teams. The procedure complied with the appropriate Institutional Review Board (IRB), and each subject provided written consent prior to cap application. For each day of testing, 3 subjects were set up with the dry sensor interface (average set-up time of less than 5 minutes), with one subject each day repeating the session as part of a wet/dry comparison. Each subject was run through a battery of tests that took approximately 90 minutes from set-up to break down. Session recordings included the following tasks: a Baseline session of Eyes Open (EO), Eyes Closed (EC), Eye Blinks, and EMG; SSVEP at 5, 10, and 15 Hz; a Baseline SSVEP EO for 2 minutes; a Rapid Serial Visual Presentation (RSVP) Video task; an RSVP Image task (with Novelty Image) and Evoked Response Potential (ERP) Task; and an Audio ERP task. As a follow-up to the live performance requirement, 2 additional subjects (2 male; ages 22 and 25) were run through the full test battery in house, using further iterations of the semi-dry interface. Changes are discussed in the Materials section, and the results and outputs were comparable to those of the earlier 9 subjects. For purposes of this paper, the results will focus on data from the RSVP Image Task (with Novelty Image) n =11. Future papers will discuss the results of other tasks performed.

The experimental task for RSVP presented 25 sets of 50 images. Each set of images comprised 49 terrain images and 1 novelty Mickey Mouse image. Use of the

novelty image elicited pronounced ERPs even in shorter test sessions. Images remained on screen for 0.2 seconds each, for a total duration of 10 seconds per set. The user was provided a 2 second pause between each set of 50 images. After every 5 sets, a longer 10 second break was provided to the subject. Users were instructed to use pauses for resting and/or blinking eyes to help minimize artifacts during testing.

To obtain ERP measures, the EEG was visually inspected for artifacts and data containing muscle artifacts or eye blinks were excluded from analysis.

To accommodate differences across dry electrode teams, the comparison between wet and dry sensor types required the following set-up. All teams recorded dry sensor data from F3, F4, P3, and P4, in addition to their remaining sensor sites (which differed between teams), while simultaneously recording wet data from F5, F1, F6, F2, P5, P1, P6, and P2. For the ABM team, this entailed removing existing sensor sites at F1, F2, P1 and P2 to accommodate the wet sensor set-up. The wet sensor data was then used to create derived F3, F4, P3, and P4 channels. Differential channels F3P3 and F4P4 were then calculated and compared between the wet and dry. For the purpose of this paper, P3 and P4 were used to show the PSD correlations from the baseline EO and EC tasks and to look at Target and Non-Target ERPs from the RSVP with Novelty.

2.2 Materials

Data for all of the ABM team's dry studies was collected using Advanced Brain Monitoring, Inc.'s commercially available B-Alert X24 Wireless EEG Headset System sampling at 256Hz for all channels. The sensor montage used for data acquisition was developed in part under previous DARPA contracts, optimized for single trial ERP analysis. Sensor sites collected were F3, F1, Fz, F2, F4, C3, C1, Cz, C2, C4, CPz, P3, P1, Pz, P2, P4, POz, O1, Oz, and O2 according to the extended International 10-20 placement. All sites were referenced to Linked Mastoids in the wireless mode. The standard wet sensors were replaced with the semi-dry sensors for data collection. The semi-dry sensor consists of a hydrogel (i.e., water absorbing polymer) with dissolved hygroscopic ingredients/components to maintain hydration, and dissolved salts to conduct electricity ionically. Maintaining hydration ensures the salts stay dissolved and that the sensors retain lower skin-to-sensor impedances for longer periods of time. The hydrogel was polymerized around a cylinder of silverized spacer fabric attached to conductive (i.e., silverized) hook Velcro. This spacer fabric served 2 important roles: 1) a structural support for the hydrogel, and 2) a transition to the fabric strip. The strip that connected to the B-Alert X24 Wireless EEG Headset was a stretchable fabric that utilized silverized thread with an insulative, polymer coating applied via chemical vapor deposition (CVD) to each of the fabric strip layers to carry the signals from the semi-dry sensor to the hardware. The stretchable fabric strip was used across all 9 subjects during data acquisition. Two additional subjects were added to the data set after the required test run, using the standard commercial strip interfaced with the semi-dry sensor, bringing the subject total to the n=11 used for group summaries found in the Results section.

For wet recording comparisons, the g.TecUSBamp 16 channel, 24-bit digitizer was used, sampled at 256Hz. Set-up of the 8 wet EEG sensor sites and the ground and reference on the earlobes required abrasions at each site and alcohol prep. All impedances were required to be below 5 kΩ prior to recording data.

3 Results

3.1 Dry vs. Wet Comparison

The EEG signal shown in Fig. 1 is derived wet F4P4 compared to the actual dry derived F4P4 montage. The signals demonstrate the similarities between the wet and dry sensors, despite the steps involved to obtain each derived differential recording.

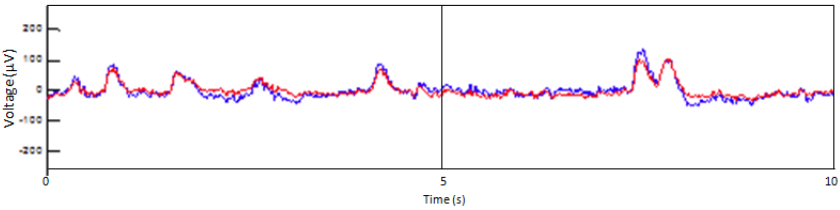


Fig. 1. Ten seconds of ABM (blue) and G.Tec. (red) EEG trace for F4P4

Subsequent comparisons between wet and dry sensors are shown using P3 and P4 to reduce the amount of data manipulation required. Figure 2 shows the PSD 1-40Hz between the dry and derived wet signal at P3 for 30 second recordings of EO, and EC recorded during the baseline session.

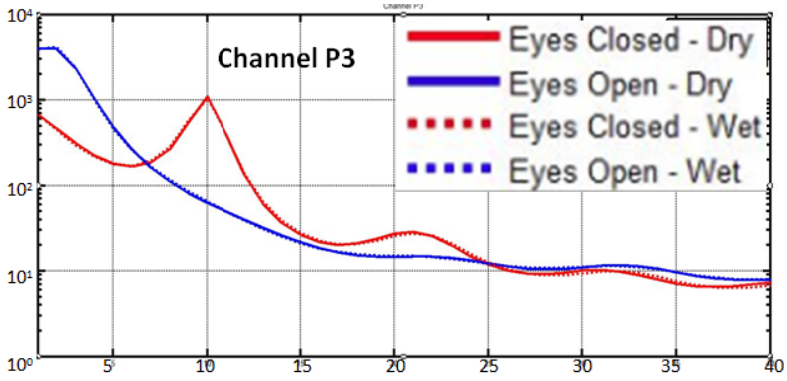


Fig. 2. 2 PSD 1-40Hz wet-dry

ERPs are shown in Figure 3 from subject 0104, recorded during the wet/dry comparison. Shown are the averaged ERPs for Target and Non Target images recorded from the dry P3 and P4 and derived wet P3 and P4 sites.

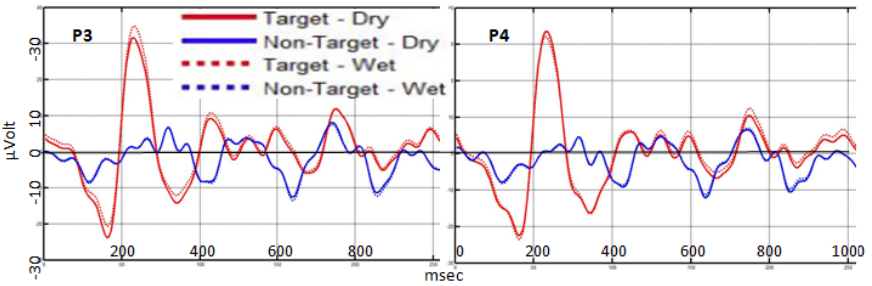


Fig. 3. Mean ERP from wet-dry comparison at P3 and P4

3.2 Grand Mean Results for Dry RSVP (n = 11)

The RSVP paradigm with Novelty was collected on 11 total subjects. All ERP graphs are shown in µVolts on the x-axis and milliseconds on the y-axis. The grand means are shown in Figure 4. We were able to obtain ERPs for all 20 sensor sites and data from 4 of the dry sensor sites are presented herein for the Mean average across subjects and the individual subject ERPs. These sites represent some of the typical ERP components across all sites. Figure 5 shows ERPs for an individual subject (ID 0102) on day one of testing, representing their first time through the test battery.

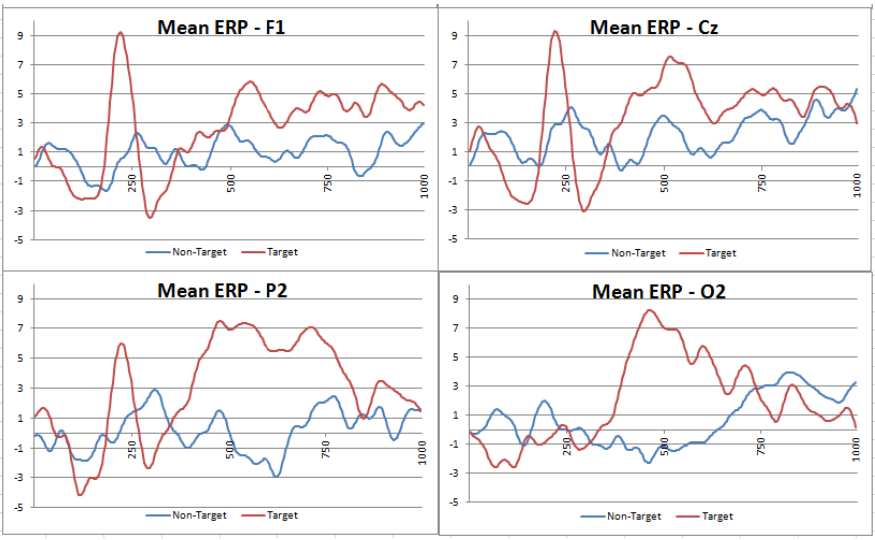


Fig. 4. Mean ERP 11 subjects Target vs. Non-Target sites F1, Cz, P2, and O2

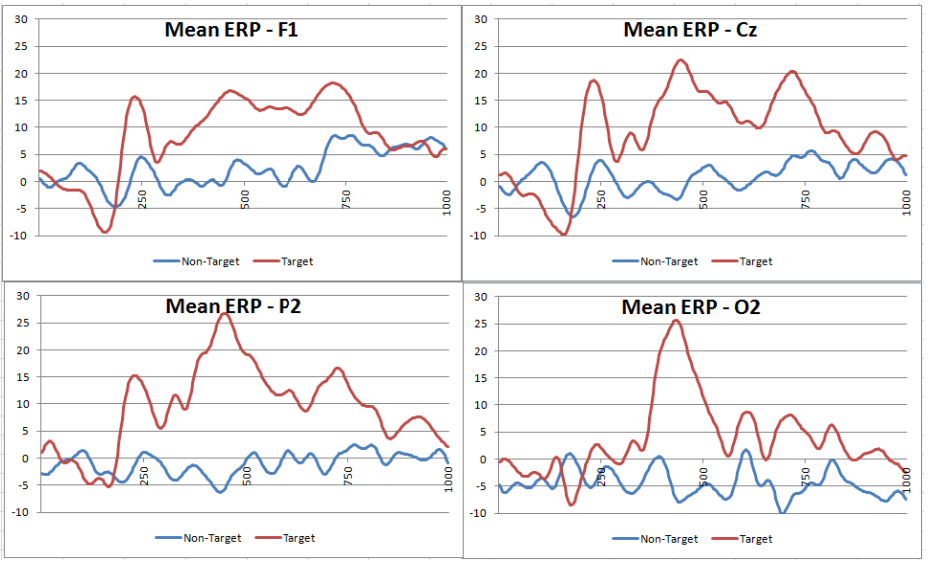


Fig. 5. Mean ERP ID0102 Target vs. Non-Target F1, Cz, P2, and O2

Figure 6 shows ERPs for individual subject (ID 0303) on day three of testing, representing their third time through the test battery. While fatigue is evident in the subject and the P300 is reduced in P2 and O2, the early P100/N200 amplitude differences still appear in F1, Cz, and P2. Figure 7 shows ERPs for another individual subject (ID 0402) on their first run through the test battery, collected after the live test run using the commercial strip interfaced with the semi-dry sensors.

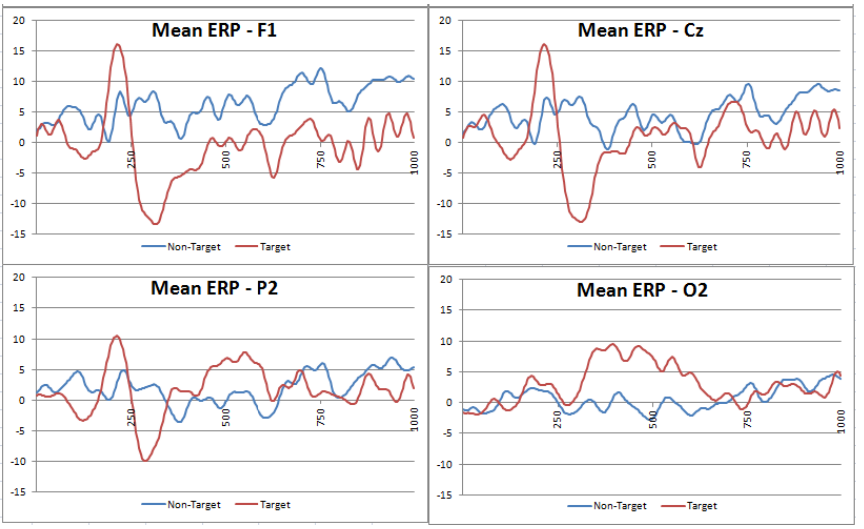


Fig. 6. Mean ERP ID 0304 Target vs. Non-Target for sites F1, Cz, P2, and O2

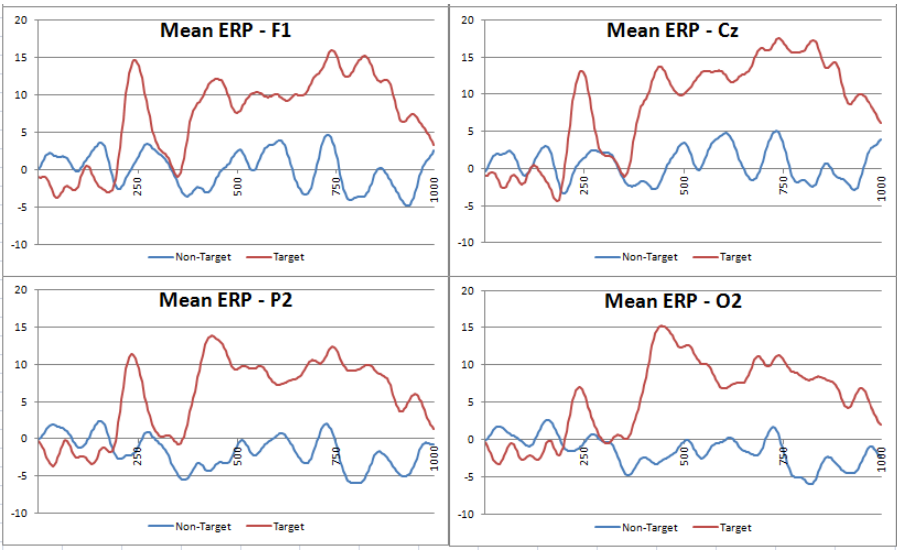


Fig. 7. Mean ERP ID 0402 Target vs. Non-Target for sites F1, Cz, P2, and O2

4 Discussion

The novel semi-dry hydrogel sensor tested in this paper proved to be a comparable alternative to the conventional (now superseded) 10/20 wet paste system. Due to the design of the all-soft semi-dry sensor, which has no rigid components, it was also able to address the main shortcomings of existing dry electrodes, providing a true real world wearable solution.

The data becomes even more compelling when considered alongside the manner in which it was collected; rather than a controlled lab wired acquisition on one or two sensor sites, EEG data was collected wirelessly as part of a live performance requirement with a 20 sensor site montage. Each day of testing included 4 sessions, each 2 hours in length, to allow for a full test battery that included set up, multiple paradigms, and break down. Once per day, one of the set-ups included a wet/dry comparison requiring real time modifications to the system to accommodate the wet sensor placements that overlapped with dry sensors. To provide common subjects across teams, all 9 subjects rotated to a new team each day with 1 subject from each team completing both a dry and a wet/dry session each day. The rotation resulted in some subjects completing 6 full sessions by the end of day 3, while the remaining subjects would each complete 3 sessions by the end of day 3.

Despite the repetition of the test battery for 6 of the 9 subjects, the Mean ERPS for Target vs. Non-Target still resulted in prominent P100, N200, and P300 features for each individual. Fatigue and repetition had the largest impact on the P300 components, but the amplitude difference between the P100 and the N200 peaks remained significant throughout testing as seen in Figure 6, iteration three for the subject.

After the completion of the live performance requirement, ABM conducted additional test runs on further iterations of the dry strip interface. While the stretchable conductive fabric used for the first 9 subjects shows great promise, the current manufacturing expenses may prove cost-prohibitive. Testing following the live performance was conducted on alternative applications of the current commercial strip that will allow rapid sensor change outs from a prepackaged form factor. The new packaging works with both the semi-dry hydrogel sensor and the currently used easy-to-apply foam and synapse cream sensor. The ERPs collected from subsequent tests with the new strip show the same ERP components from the live performance test on the group of 9. Ongoing additional testing continues to support equivalence between the two dry strip interfaces.

ABM plans to continue refining the dry sensor, ultimately arriving upon a commercially available dry sensor option that the end user can switch between depending on the goals and applications of the intended study.

Acknowledgements. This work was supported by The Defense Advanced Research Projects Agency (government contract number NBCHC090054). The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for Public Release, Distribution Unlimited.

References

1. Berger, H.: uber das Elektroenkephalogramm des Menschen. Eur. Arch. Psychiatry Clin. Neurosci. 87, 527–570
2. Guger, C., et al.: Comparison of dry and gel based electrodes for P300 brain-computer interfaces. Front. Neurosci., doi:10.3389/fnins.2012.00060
3. Wang, L., et al.: PDMS-Based Low Cost Flexible Dry Electrode for Long-Term EEG Measurement. IEEE Sensors Journal 12(9) (September 2012)
4. Slater, J., et al.: Quality Assessment of Electroencephalography Obtained From a "Dry Electrode" system. Journal of Neuroscience Methods 208, 134–137 (2012)
5. Forvi, E., et al.: Preliminary Technological Assessment of Microneedles-Base Dry Electrodes for Biopotential Monitoring In Clinical Examinations. Sensors and Actuators A 180, 177–186 (2012)
6. Dias, N.S., et al.: Wireless Instrumentation System Based on Dry Electrodes for Acquiring EEG Signals. Medical Engineering & Physics 34, 972–981 (2012)
7. Ghoshdastider, U., et al.: Development of a Wearable and Wireless, Modular, Multichannel, EEG-System Utilising Dry-Electrodes for Long Time Monitoring. Biomed Tech. (2012), doi:10.1515/bmt-2012-4056
8. Berka, C., et al.: Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset. International Journal of Human-Computer Interaction 17(2), 151–170 (2004)
9. Stevens, R., et al.: Modeling the Neurodynamic Complexity of Submarine Navigation Teams. Computational and Mathematical Organization Theory (2012)

10. Berka, C., et al.: Accelerating Training Using Interactive Neuro-Educational Technologies: Applications to Archery, Golf, and Rifle Marksmanship. International Journal of Sports and Society 1(4), 87–104
11. Chung, J.W., et al.: Treatment Outcomes of Mandibular Advancement Device for Obstructive Sleep Apnea Syndrome. Chest 140, 1511–1516
12. Westbrook, P., et al.: Description and Validation of the Apnea Risk Evaluation System: a Novel Method to Diagnose Sleep Apnea-Hypopnea in the Home. Chest 128, 2166–2175
13. Behneman, A., et al.: Neurotchnology to Accelerate Learning. NEST (2012) (in Press)

Real-Time Workload Assessment as a Foundation for Human Performance Augmentation

Kevin Durkee[1], Alexandra Geyer[1], Scott Pappada[1], Andres Ortiz[1], and Scott Galster[2]

[1] Aptima, Inc., USA
[2] Air Force Research Laboratory, USA
{kdurkee,ageyer,spappada,aortiz}@aptima.com,
scott.galster@wpafb.af.mil

Abstract. While current military systems are functionally capable of adaptively aiding human operators, the effectiveness of this capability depends on the availability of timely, reliable assessments of operator states to determine when and how to augment effectively. This paper describes a response to the technical challenges associated with establishing a foundation for reliable and effective adaptive aiding technologies. The central component of this approach is a real-time, model-based classifier and predictor of operator state on a continuous high resolution (0-100) scale. Using operator workload as a test case, our approach incorporates novel methods of integrating physiological, behavioral, and contextual factors for added precision and reliability. Preliminary research conducted in the Air Force Multi Attribute Task Battery (AF_MATB) illustrates the added value of contextual and behavioral data for physiological-derived workload estimates, as well as promising trends in the classification accuracy of our approach as the basis for employing adaptive aiding strategies.

Keywords: Workload, Augmentation, Human Performance, Modeling and Simulation, Physiological Measurement.

1 Introduction

To address the modern threat environment, military operations must overcome a variety of demands and resource constraints, such as manpower limitations, information overload, sustained long-term missions, and an increasingly complex decision space. This reality leads to our military force being more vulnerable to performance decrements related to increases in cognitive workload, stress, and fatigue. There are available technological solutions that could help mitigate these types of performance decrements through adaptive aiding and, consequently, benefit the effectiveness of active operational systems. Traditional approaches to designing user interfaces (UI) typically result in a fixed presentation of information throughout the entirety of the operator interaction with a control station; however, human operator states (e.g., workload, engagement, and affect) are dynamic. For instance, if the system detects that the human operator is experiencing high workload, as when an remotely piloted aircraft (RPA) pilot must monitor a noisy video feed of a crowded marketplace while

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 279–288, 2013.

simultaneously attending to frequent audio and chat communications for task-relevant information, the system could alter the interface to (1) eliminate all the irrelevant information that may clutter the display to reduce the workload demand, and (2) bring into focus central information that needs attention [1]. While different operator states often entail different ideal interface configurations, traditional approaches to UI cannot accommodate this demand.

The feasibility and overall effectiveness of adaptive performance augmentation is dependent on timely and reliable assessments of a human operator's state. The ability to accurately and autonomously define an operator's state, particularly in real-time, has been a much desired yet difficult to achieve capability that has hindered the ability to employ adaptive aiding technologies. One approach that has generated much interest in recent years focuses on the use of physiological data to classify an operator's state. Previous research has shown that physiological measures can be used to detect operator state [2, 3]. Recent improvements in reliability, level of invasiveness, set-up time, and cost of physiological measurement makes it even more compelling. Physiological data also serves as an objective source of information and is theoretically available from any person working in any domain, in contrast to behavioral and situational data which are likely to vary greatly across different work environments.

However, from the perspective of developing an operationally deployable capability for estimating operator states, there have been limitations with regard to: (a) the ability to produce a model with high levels of accuracy across individuals, particularly when the operator state model has not been "trained" to a specific individual; (b) the ability to derive an accurate classification from available real-time data, as opposed to post-hoc analysis in which a much larger spectrum of data are available (e.g., future events, subjective responses, etc.); and (c) the ability to pinpoint the operator's state with high resolution and update frequency. Some of the most successful operator state classification efforts to date have made progress in this endeavor by collecting large sums of data from a specific individual, and subsequently training a custom operator state model for that same individual with machine learning based methods [4]. While this work produced invaluable insights on the possibilities of operator state classification, there are practical limitations to shaping specially trained models to each individual operator using a particular system. More recent work has started to explore cross-subject workload classification [5], however this body of research remains in the early stages. In addition, a prominent theme in the literature to date is the classification of operator states according to very discrete categories, such as "low workload" and "high workload", as well as outputting these categorical state estimates at infrequent intervals. When attempting to employ automated augmentation strategies, the lack of granularity allotted by a "low vs. high" classification and at infrequent update rates may prevent a system from tracking the necessary detailed trends and subtle fluctuations over time that can greatly affect the operator's need for intervention.

In addition, the ideal adaptive augmentation system would be able to incorporate predictions of operator state and its expected impact on human performance. Predictive capabilities would provide an invaluable tool for proactively address problems before they occur. Unfortunately, operator state predictions have not been thoroughly explored, as much of the published research has been focused on historical and

real-time diagnosis of operator state. These predictive capabilities are also held back by the lack of a reliable, continuous, and frequently updated estimate of operator state that supplies the required level of granularity and volume of data necessary to make quality predictions. Collectively, these gaps illustrate the need for a forward-looking approach that can establish an extensible foundation for adaptive aiding strategies; one that is both practical for application and improves the likelihood that dynamic interventions will have a beneficial effect on operator state and job performance.

2 Approach

The objective of our research is to expand upon this existing foundation of research to identify the most relevant and sensitive multi-modal measures of operator states (i.e., neural, physiological, behavioral) and develop algorithms that can assess these states in real time for the purpose of enabling various performance augmentation strategies. In response to this technical challenge, we have designed and implemented an approach that intends to lay a foundation for adaptive aiding technologies to be transitioned to operational system usage.

Our approach relies on innovative physiological-based operator state modeling and classification techniques being formulated and tested within the Air Force Research Laboratory's (AFRL) "Sense, Assess, Augment" taxonomy [6]. To fulfill the "Sense" component of this framework, we have developed a flexible architecture (Figure 1) for collecting and processing physiological, behavioral, and situational data from disparate sources in real-time into a centralized location. The "Assess" component of this framework employs a machine learning based modeling approach that is trained from data sets spanning four categories: Physiological, Self-reported factors, Performance, and Situational. As our test case, the current focus of the assessment component is on operator workload classification as a function of these four categories, given that workload has a demonstrated relationship to task performance and thus is an "augmentable" construct. Lastly, the "Augment" component seeks to "close the loop" on sustained human performance by leveraging the accessibility of real-time continuous workload estimates as the basis for when and how to aid performance. For the purpose of this paper, we focus primarily on the "Sense and Assess" portions of this framework as a stepping stone to achieving the end goal of effective real-time adaptive augmentation strategies.

Our modeling approach is unique on several fronts. First, the inclusion of expansive contextual information to support the model's ability to interpret noisy physiological data has not been substantially explored by other published approaches. We theorize that data characterizing an individual's antecedent health and lifestyle factors, real-time task performance, and situational data from the task environment provide beneficial insight into why physiological patterns occur, thus supporting the ability to "sift through the noise" and ultimately obtain the most meaningful data for operator state classification.

Second, this approach supplies a real-time output with a continuous high-resolution (0-100) scale. We accomplish this by applying machine learning methods to train a

model that identifies the best fit between these available real-time data sources and subjective operator state measurements collected from our experimental paradigm (described in the next section). With respect to our model training approach, we inject noise into each subjective measurement for each corresponding trial to generate an operator state estimate along a continuous scale for model training, under the assumption that few, if any, meaningful operator states are perfectly static over time. Because it is impractical, if not impossible, to obtain operator responses at very frequent intervals (e.g., once per five seconds), it is important to rely on a theoretically-grounded relationship between an available, measurable factor (or set of factors) and the modeled construct of interest as the basis for incorporating noise. The complexity of this component of our approach can range from simple to highly complex depending on the modeled construct and tolerance to error. As an example, for our test case of modeling operator workload, we add noise to self-reported workload ratings based on specially designed algorithms that process contextual data about the situation at each point in time to produce the direction and magnitude of noise.

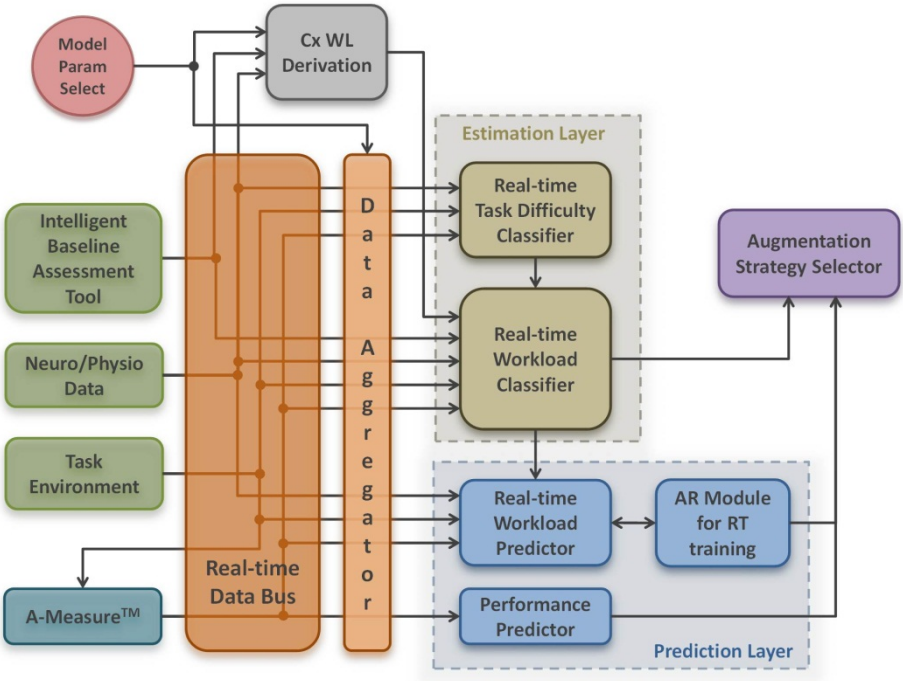


Fig. 1. Data aggregation and modeling architecture for operator state assessment as a foundation for automated adaptive aiding strategies

Third, this approach employs on-line model training capable of improving the precision of the operator state estimation across different individuals over time. The on-line training process is triggered by a scalable set of operator-driven inputs dependent on the state being modeled. Ideally, the scale of these inputs is set to consume the

lowest possible time, effort, attention, and frequency of input from the operator (e.g., 5-second response per every 30 minutes). Using this trigger, a set of sub-components within our architecture dynamically updates the model weights using the operator input data in conjunction with recent physiological, behavioral, and situational data that has occurred during a corresponding timeframe, resulting in more accurate and individualized estimate of operator state that improves over time without the need for a priori custom-built classifiers for each human operator.

Lastly, the predictive layer of this approach utilizes memory of historical data to help facilitate informed, and proactive, augmentation decisions based on expected operator state and performance. The predictive accuracy is, as one would expect, dependent on the level of granularity and update frequency of the real-time operator state classifier. For example, workload estimates on a 0-100 scale and updated once per every five seconds allows a trained model to monitor subtle trends and changes not otherwise possible with highly discrete classifiers (e.g., high versus low); this may potentially be the difference between knowing when, and when not, to intervene with an augmentation strategy. In addition, forecasted knowledge of the situation – such as when a highly tactical and attention-demanding phase of a mission is known to occur – is valuable, if not essential, context that adds to the accuracy of workload and performance predictions.

3 Current Study

3.1 Overview

To develop a prototype operator state model based on this approach, we conducted a model training study at AFRL's Human Universal Measurement and Assessment Network (HUMAN) Laboratory. Our primary objective was to generate data sets that would allow an operator state model of workload to be trained within our defined technical approach. For the scope of this paper, our reporting focuses primarily on model classification accuracy in relation to related published work. Secondary objectives of this study were to validate that subjective workload ratings to be used for training the workload model indeed correspond to the intended task difficulty, and conduct exploratory analysis on the degree to which workload fluctuations correspond to performance fluctuations. These latter objectives are important as a preface to our future research on developing effective augmentation strategies.

3.2 Task Environment

The task environment for this study was based upon a modified version of the Air Force Multi-Attribute Task Battery (AF_MATB) [7]. This PC-based aviation simulation requires an operator to perform an unstable tracking task while simultaneously monitoring warning lights and dials, responding to simulation-generated auditory requests to adjust radio frequencies, and managing simulated fuel flow rates using various key presses. Our rationale for using this task environment was threefold. First, MATB has been used as a testbed to train and develop other models of operator

workload [5], which provides our approach with a benchmark for comparison. Second, MATB allows for linear titration of workload on a high-resolution scale, which provides the necessary task conditions to model beyond "low versus high workload" prior to injecting noise. Third, MATB has long and rich history of research findings that provide a deep understanding of how each task module affects operator workload, as well as the interactions between these factors.

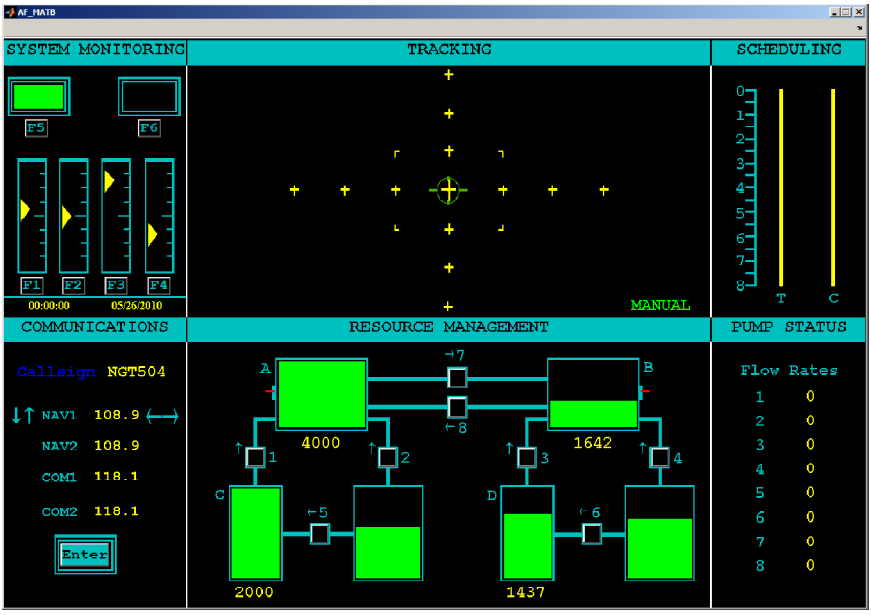


Fig. 2. The operator interface for the AF_MATB task

3.3 Participants

Ten participants served as operators of the AF_MATB system for this study. The only requirement for participation was a familiarity with computer-based systems. Seven participants were male and three participants were female. The age of participants ranged from 23 to 47 years old, with a mean age of 32 years old.

3.4 Experimental Design

Task difficulty was the only independent variable (IV) for this study. We selected task difficulty because this manipulation has been an effective method for inducing varying levels of operator workload [8]. In our attempt to obtain the highest possible level of model granularity, we relied upon 15 levels of task difficulty that intend to linearly span the full range of workload (i.e., low to high). Accordingly, this study employed a one-way experimental design in which a single IV (task difficulty) was manipulated across 15 conditions in order to assess its effects on each dependent variable (DV).

All participants experienced the same 15 conditions; however, the sequence of conditions was counterbalanced to mitigate order effects.

3.5 Dependent Variables

The dependent variables (DVs) were physiological, self-reported, and performance measures collected from participants. Physiological measures included EEG, ECG, and eye-tracking activity (e.g., pupil diameter, fixations, blinks, etc.). Self-reported measures included antecedent lifestyle factors (e.g., demographics, level of exercise, video game experience), recent behavioral factors that can affect physiological state (e.g., sleep quality, current sleepiness, caffeine and food intake), and subjective workload assessments of each condition as measured via the NASA Task Load Index (TLX) scale [9]. Performance measures included primary task performance on the AF_MATB tracking task (distance from centerpoint) and secondary task performance on the lights/gauges task (response time and accuracy).

3.6 Procedures

Each participant went through two sessions: training and data collection. During the training session, participants acquired hands-on training by operating the system during practice scenarios ranging across easy, medium, and hard difficulty conditions. Our goal was to eliminate learning effects during the data collection phase to the extent possible. For the data collection session, participants operated the AF_MATB environment through 15 five-minute scenarios while being monitored with physiological sensors and behavioral data capture software. Each of the 15 scenarios varied by task difficulty and was presented in a quasi-randomized order with five blocks of three scenarios per block. The three blocks in each scenario consisted of a low, medium, and high difficulty block. Physiological sensors collected data on eye movements, blinks, pupil diameter, EEG, and ECG. At the end of each trial, participants completed the NASA TLX questionnaire provided electronically on the AF_MATB.

4 Results and Discussion

4.1 Model Training Results

Within the scope of this study, there are several ways to evaluate the utility of trained model results. First, we evaluated the absolute error (expressed as mean absolute difference percent) between the model's output and the reference continuous workload estimator values upon which the model was trained, which came to an average of 35% for all participants across all trials. For some participants, the average error across trials reached as low as 15%, although other participants produced greater than 50% error. We concluded that while we may have collected many valuable inputs that account for the majority of workload variance for specific individuals, there could be individual differences that were not sufficiently measured.

Second, we analyzed classification accuracy of the trained model when applied retroactively to participant data without providing the model with any direct workload-related input. While categorization is not the ultimate goal of this approach, it is useful as a means for comparing this work to known benchmarks in the literature. Using classification accuracy for low versus high workload, the prototype model produced mean 82.7% accuracy when averaged for entire trials, and 75.7% accuracy on a per five-second basis. We also went a step further by randomly removing two participants from the training set and applying the adjusted model to these removed participants. When averaged for entire trials, the adjusted model produced a mean 87.5% accuracy for low versus high classification for these two participants, and 77.8% on a per five-second basis. When considering our use of continuous high-resolution output as the basis of these classifications – as well as the small sample size and our inclusion of outliers – these results appear to compare favorably to similar work [4, 5]. In addition, our preliminary analysis on the benefits of on-line model training techniques (which are not reported here due to intended scope) has revealed promising trends with regard to additional accuracy generated due to dynamic model weight adjustments over time based on the individual performer.

While the per five-second classification accuracy of our workload model is difficult to empirically validate at this time (i.e., it is not feasible to obtain self-report data every five seconds for comparison), these results provide a quality baseline standard from which to expand our forthcoming work. Our future research will: (a) quantify the benefit of a larger sample size and on-line model training; and (b) identify methods to validate high-resolution output of our approach beyond categorical levels.

4.2 Secondary Analyses

Secondary objectives of this study were to validate that subjective workload ratings to be used for training the workload model indeed correspond to the intended task difficulty, and conduct exploratory analysis on the degree to which workload fluctuations correspond to performance fluctuations. While these findings are not directly related to the formulation of our operator state modeling approach, they can be used as a preface to our future research on developing effective augmentation strategies.

Correlation between participants' self-reported NASA TLX ratings and intended experimental difficulty (1-15) was approximately 0.67, demonstrating that workload was indeed reasonably well connected to the intended task difficulties of scenarios. We further validated this assumption by grouping continuous workload measures used for model training based on intended task difficulty/workload: Low (difficulties 1-5), Medium (difficulties 6-10), and High (difficulties 11-15). Based on these groupings, there was a statistically significant difference between mean continuous workload measures used for model training across each of three groups ($p<0.0001$). Furthermore, we analyzed the addition of our noise injection algorithm to the NASA TLX responses to generate the continuous workload estimates for model training. When averaging the resulting continuous workload estimates across trials, we obtained a correlation of $r = 0.99$ with the actual reported NASA TLX values, which demonstrated the noise injection algorithm did not overly skew workload responses.

Lastly, at an exploratory level we investigated the degree to which the model's estimates of workload provided identifiable clues to when performance decrements might occur. This was an informal analysis done to obtain a realistic expectation as to how frequently performance decrements could be identified proactively, using workload as the "leading indicator" and/or "trailing indicator" of their occurrence. The example illustration in Figure 3 demonstrates one recurring trend in which a performance decrement can serve as a leading indicator of workload spikes, followed by subsequent behavioral changes in reaction to these effects. Currently, we are quantitatively formalizing the complex relationships between workload and performance as a precursor to intelligent augmentation strategy selection in real-time mission settings.

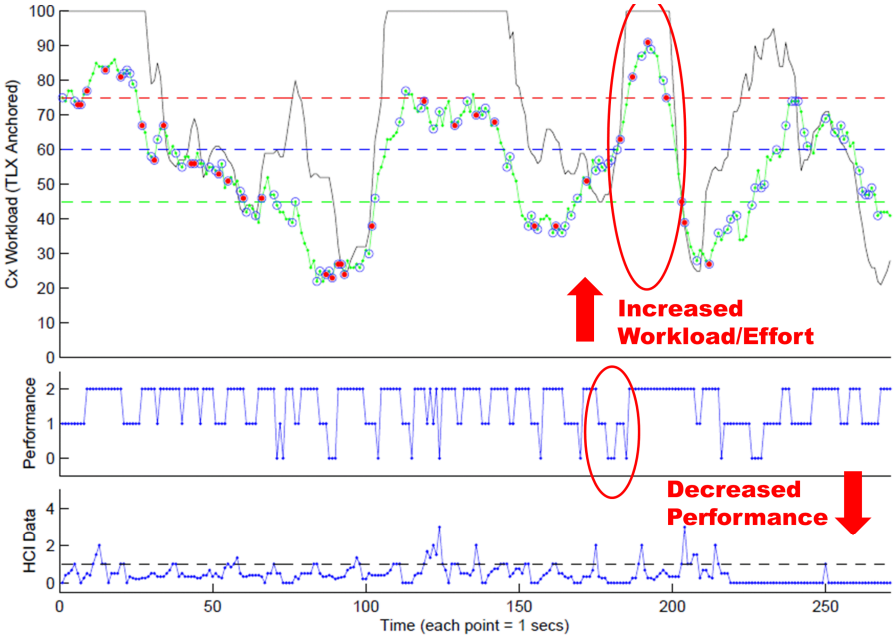


Fig. 3. Example of a workload spike as a leading indicator for a performance decrement

5 Conclusions

This paper described a novel technical approach for establishing real-time estimates of operator states on a continuous, high-resolution scale for the purpose of improving the ability to employ effective adaptive aiding strategies for performance augmentation. Using operator workload as a test case, our research to date has served as a key stepping stone with regard to establishing a level of accuracy in line with the published state of the art. Future research will focus on improving model accuracy through additional data collection, optimizing components of the model architecture (e.g., on-line training), and additional measures that may account for a larger percentage of workload variance. A critical next step is also the design of a model validation

paradigm that enables empirical investigation of workload estimation accuracy on a continuous 0-100 scale. Finally, we will quantitatively represent the complex relationships between workload and performance, which may provide substantial benefit to the employment of automated aiding strategies to mitigate performance decrements.

Acknowledgement. This material is based upon work supported by the Air Force Research Laboratory (AFRL) under Contract No. FA8650-11-C-6236. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL. The authors would like to thank: Seamus Sullivan, Noah DePriest and Zachary Zuzack for software support; Matt Middendorf, Michael Hoepf, Cassandra Christman, and Chelsey Credlebaugh for data collection support; and Justin Estepp for feedback on our technical approach.

Distribution A: Approved for public release; distribution unlimited. 88ABW Cleared 4/02/2013; 88ABW-2013-1591.

References

1. Parasuraman, R.: Neuroergonomics: Brain, cognition, and performance at work. Current Directions in Psychological Science 20, 181–186 (2011)
2. Wilson, G.F., Eggemeier, F.T.: Physiological measures of workload in multi-task environments. In: Damos, D. (ed.) Multiple-task Performance, pp. 329–360. Taylor & Francis, London (1991)
3. Schnell, T., Keller, M., Macuda, T.: Application of the Cognitive Avionics Tool Set (CATS) in Airborne Operator State Classification. In: Augmented Cognition International Conference, Baltimore, MD (2007)
4. Wilson, G.F., Russell, C.A.: Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. Human Factors 45(4), 635–644 (2003)
5. Wang, Z., Hope, R.M., Wang, Z., Ji, Q., Gray, W.: Cross-subject workload classification with a hierarchical Bayes model. NeuroImage 59(1), 64–69 (2012)
6. Galster, S.: Sense-Assess-Augment: A Taxonomy for Human Effectiveness. In: Proceedings of the Seventeenth International Symposium on Aviation Psychology. Dayton, OH (in press)
7. Miller, W.D.: The U.S. Air Force-developed adaptation of the Multi-Attribute Task Battery for the assessment of human operator workload and strategic behavior (Tech. Rep. No. AFRL-RH-WP-TR-2010-0133) (2010)
8. Backs, R.W., Seljos, K.A.: Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working memory task. International Journal of Psychophysiology 16, 57–68 (1994)
9. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Peter, A.H., Najmedin, M. (eds.) Advances in Psychology, vol. 52, pp. 139–183. North-Holland, Amsterdam (1988)

Using the EEG Error Potential to Identify Interface Design Flaws

Jeff Escalante[1], Serena Butcher[1], Mark R. Costa[2], and Leanne M. Hirshfield[3]

[1] Hamilton College and
[2] School of Information Studies and
[3] S.I. Newhouse School of Public Communications, Syracuse University
escalantejeff@gmail.com, sbutcher@hamilton.edu,
{mrcosta,lmhirshf}@syr.edu

Abstract. There are a number of limitations to existing usability testing methods, including surveys, interviews, talk-alouds, and participant observations. These limitations include subject bias, poor recall, and inability to capture fleeting events, such as when a UI functions or behaves in a manner that contradicts user expectations. One possible solution to these problems is to use electrophysiological indicators to monitor user interaction with the UI. We propose using event related potentials (ERP), and the error potential (ErrP) more specifically, to capture moment-to-moment interactions that lead to violations in user expectations. An ERP is a response generated in the brain to stimuli, while the ErrP is a more specific signal shown to be elicited by subject error. In this experiment we monitored subjects using a 10-channel electroencephalogram (EEG) as they completed a range of simple web browsing tasks. However, roughly 1/3 of the time subjects were confronted with poor UI design features (e.g., broken links). We then used statistical and machine learning techniques to classify the data and found that we were able to accurately identify the presence of error potentials. Furthermore, the ErrP was present when the subjects encountered a UI design flaw, but only during the more 'overt' examples of our design flaws. Results support our hypothesis that ERPs and ErrPs, can be used to identify UI design flaws for a variety of systems, from web sites to video games.

Keywords: EEG, usability testing, error potential.

1 Introduction

Usability testing is a critical part of the design process, and can be conducted using a number of different methods. Surveys, talk-alouds, focus groups, and interviews all offer ways to explore the user experience. However, all of these methods fail to adequately capture fleeting interactions by the users; post-use interviews rely on imperfect memory to create a narrative summary of the experience, and talk-alouds interrupt the natural flow of use. Furthermore, these micro-events can escape the conscious processing of the user yet can have a cumulative negative impact on users' satisfaction (Hirshfield et al., 2013). Examples of these micro-events include

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 289–298, 2013.

interacting with broken links, improperly operating forms, buttons working incorrectly or performing unexpected operations, users clicking on the wrong button because it is too close to another desired button, etc. Additionally, capturing these events with notes during an observation is slow and prone to error, and it is also possible for users to correct their action quickly enough that the observer and the user fail to notice.

In addition to technical errors, users may also experience frustration when working with an interface that violates customs or conventions. Well-designed UIs should not violate users' expectations—that is—the action items available to a user working with a computer system, and the feedback provided after the user takes an 'action', should fit with the users' expectations—enabling them to immerse themselves with the task at hand. As an illustrative example: the CTRL+Left Click combo is for selecting multiple items in a list; assigning the delete function to this key combo would violate user expectations.

Identifying these types of events is a well-studied problem in the human factors domain. When usability experts want to gather information that extends beyond that gleaned through accuracy and speed data, they often depend on qualitative feedback gathered via surveys, focus groups, and interviews. This can cause problems, as it has has been demonstrated that self-report measures can be very unreliable (Shneiderman & Plaisant, 2005) in that they often include subject biases, they lack real-time information about user experiences throughout a task, and they are limited by many users' inability to accurately describe their experience while working with a given UI design. In the current study, we look to neuroscience methods to attempt to provide additional, and valuable, quantitative information throughout a usability study that can serve as a real-time indicator that a given user's expectations with the UI have been violated.

Event Related Potentials (ERPs) are electrophysiological responses in the brain to an internal or external stimulus. ERPs can be found in EEG data as specific waveforms preceding or following a stimulus. ERPs have several components; for this study we focus on the 'Error Potential' (ErrP). The ErrP is a signal which has been shown to be elicited by subject error and is present within milliseconds after a person realizes that he or she has made a mistake (via immediate feedback presented to them that indicates they were incorrect). Results from recent studies indicate that the ErrP can also be elicited by more general violations in an individual's expectancy (Oliveira et al., 2007). This study attempts to take state-of-the-art error potential research a step further, applying it to the concept of usability testing. The potential of using the non-invasive EEG device during human-computer interactions has been proposed, and validated, by a wealth of recent research (Tan and Nijholt, 2010). The EEG is lightweight, portable, and it has been implemented wirelessly, allowing the monitoring of computer user's brain activity during real world settings.

In this study we test the hypothesis that the ErrP will appear in the EEG data after an interaction with common interface design mistakes. Additionally, we test the feasibility of using machine learning to accurately classify these interactions.

2 Background and Literature Review

An EEG measures the field potentials produced by the firing of neurons in the brain (Tatum, Husain, & Benbadis, 2008). EEG devices have channels, corresponding to the number of electrodes used to capture the data. The equipment used for this study is an ABM 10-channel wireless EEG (http://www. advancedbrainmonitoring.com). One channel is assigned to the reference electrodes, commonly affixed to the mastoid bones behind each ear of the subject. An integrated voltmeter records the difference in voltage between the site of interest and the reference electrode. This reference site is chosen as to be relatively uninfluenced by activity in the area of interest (Coles & Rugg, 1996).

EEGs can determine when activity occurs in the brain, although most are unable to distinguish beyond a gross estimate where the activity occurred. There are some exceptions to this rule for EEGs with a high number of electrodes (128). A 10 channel EEG is relatively easy to set up, but offers limited spatial resolution. In exchange for that limitation, the EEG offers high temporal precision, with a sampling rate of 256Hz (~4ms). EEG also has a few other advantages over other methods of recording brain activity. For example, EEG recordings are less sensitive to a subject's movement than fMRI. EEG devices are also small and relatively portable, and are therefore much more suitable for simulating natural human conditions and surroundings.

While EEG produces a number of usable data streams, in this study we focus on one, the event-related potential (ERP) – a spike is generated in response to an internal or external stimulus. A common way of measuring ERPs is to "time-lock" a stimulus to the EEG signal. For example, a time period could be defined that extends .5 sec before the onset of the stimulus and ends 1 sec after. Within this time period, there may be changes in the brain's electrical activity that relate specifically to the stimulus.

2.1 ERP

The ERP is a response in the brain to an internal or external stimulus, and appears in the EEG data as specific waveforms preceding or following a stimulus, after filtering the data. There are a number of components in an ERP; for this study we will confine our analysis to the 'Error Potential' (ErrP). The ErrP in turn has two sub-components error-related negativity (ERN or Ne) and error-related positivity (Pe). In addition, modifications of the signal have been discovered, particularly feedback error-related negativity (fERN), response error-related negativity (rERN), and the interaction ErrP.

Researchers discovered the Error-related negativity (ERN) while observing subjects committing errors in simple choice response tasks. The signal "takes the form of a sharp, negative-going deflection of up to 10 μV in amplitude, and is largest at electrodes places over the front and middle of the scalp" (Gerhring, Coles, Meyer, & Donchin). Additionally, the signal begins immediately after the incorrect response, and peaks 80 - 150ms later (Gehring 1993).

A Pe, characterized by a positive deflection in the signal, can follow an ERN. The Pe occurs when the subject becomes aware of the error (Nieuwenhuis et. al., 2001), but is not dependent upon error correction (Falkenstein et. al. 2000). If a Pe appears, it

immediately follows the ERN, occurring 200-500ms after the incorrect response (Falkenstein et. al., 2000). ErrPs can also be elicited in subjects by giving negative feedback (Miltner et. al., 1997). When both types of ErrPs are involved, the standard response referred to as rERN (response error-related negativity) and the feedback-elicited response is referred to as fERN (feedback error-related negativity).

Few studies have examined the error potential in relation to human-computer interaction. In one study, Ferrez and Milán (2005) found that ErrPs were elicited by incorrect interpretation of a subject's intent by a computer interface, which they dubbed an "interaction ErrP". The interaction ErrP is similar to an rERN, occurring immediately after the stimulus and taking the same shape. However, the interaction ErrP has a sharper negative peak and broader positive peak. The goal of this study is to build on Ferrez and Millan's BCI research, demonstrating how Interaction ErrPs can be detected and used to evaluate human-computer interactions for non-disabled users.

2.2 Processing the EEG Data

EEG signal captures the data of thousands of ongoing processes in addition to the response generated by the stimulus. Consequently, it would be difficult to detect the activity related to the stimulus after just one trial. The averaging method is often used to overcome this limitation. Averaging involves generating a set of values by recording a number of different time-locked trials for the same event, then averaging the set to produce a value for that type of event (Coles & Rugg, 1996; Mouraux & Iannetti, 2008).

Averaging, although popular, does have certain shortcomings. One problem is that, due to the averaging between many trials, this procedure can not directly measure the ERP elicited by an individual event. Because of this, the resulting data must be analyzed on its own and not compared to other measures such as reaction time for an individual stimulus. There is in addition a problem with processing when the waveform of a trial has a bimodal distribution (two different modes with distinct peaks). In this case, the average amplitude will not correspond to the actual amplitude of any of the trials (Coles & Rugg, 1996).

Ferrez and Millan (2005) used machine learning techniques to create a statistical Gaussian classifier to predict, on a single trial basis, whether or not a portion of EEG data indicated a correct, or an incorrect subject response. Similarly, Hirshfield et. al. (2009) implemented signal processing and machine learning algorithms to conduct single trial analyses on their EEG data. They split the continuous EEG data into small 2 second windows, with windows overlapping every second, and then took a Fourier transform of the data in each window. For each window, they computed the magnitude and phase of the signal and the spectral power of the signal in the delta (1-4Hz) theta (4-8Hz), alpha (8-12Hz), beta-low (12-20Hz), beta-high (20-30Hz), and gamma (30-50Hz) frequency bands. They also computed the coherence and cross spectrum between each channel for each frequency band in each window. This resulted in over 6,200 features for each instance. They then use blocked cross validation to select most

relevant attributes for single trial classification. They used an information gain heuristic followed by Weka's CfsSubsetEval function to choose the features that best predict the class label in the training data. We will be using similar machine learning algorithms to analyze our data on a single-trial basis in the current study.

3 Methodology

For each subject, we employed a two phase experiment. The first phase involved using a localizer task, in this case a difficult memory task which was time-locked to EEG data. Localizer tasks use validated tests to elicit, and verify our ability to capture, the ErrP. In the second phase subjects completed simple interface tasks that were programmed to have errors 1/3 of the time. Finding ErrPs time-locked to the interface errors would support our argument that the ErrP is useful for UI testing, making the process more sound, quantitative, and scientific.

Ten students from Hamilton College were recruited to take part in this study (8 males, mean age = 19.8). All subjects indicated that they had no history of mental disability and that they were healthy and prepared for the study beforehand. All subjects gave written consent and this study was approved by the institutional review board at Hamilton College. Subjects were fitted with a B-Alert X10 EEG headset before beginning the study. One of the subjects' data was discarded due to technical difficulties during measurement.

For the first part of the study, we utilized a difficult variation of the Sternberg memory task (Sternberg 1966) using E-Prime software (Psychology Software Tools, Inc.). The Sternberg task has been used repeatedly to elicit the ErrP. In the memory task, subjects must identify whether or not a probe letter is contained in a set of previously presented letters. Each subject underwent 150 trials in 15 cycles of 10 sets each. Sets were comprised of 3, 5, 7, 9, and 11 randomly selected consonants. Each set size was displayed to the user 15 times with a correct probe following, and 15 times with an incorrect probe following in total. Each cycle was preceded by a blank screen for .5 sec, and each letter was flashed for 1.2 sec, with a .5 sec blank screen in between each letter display. At the end of the list, the focus (a blue cross) was displayed for 1 sec, followed by a .5 sec blank screen, and then the probe letter, in red. After the subject's response, the next sequence would begin. Responses were recorded and time-locked to the EEG signal.

In the second part of the study, subjects went through 5 different mini-interface tasks in a randomized order. Each task contained 30 trials, 10 of which had an intentional error (error trials).

Symbols (Figure 1): A list of 5 symbols (thanks to the noun project - http://thenounproject.com) are randomly selected and presented for each trial. The subject is instructed to select the symbol which most accurately portrays a word presented above the row of symbols. The normal trials report a correct answer, and the error trials report an incorrect answer.

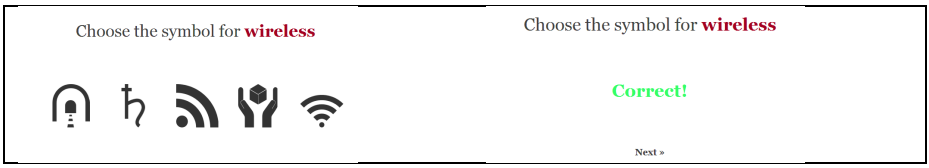


Fig. 1. Symbols

Links (Figure 2): One sentence (written in latin to prevent any comprehension or reading interference) is presented on the screen, and the subject is instructed to click on the link. One or two randomly selected words are linked, indicated by a difference in color, underline, and hover effect. In the normal trials, clicking on the link moves the user on to the next trial with no delay. In the error condition, the link text is immediately changed to "error: try again", and there is a 1.2 second delay before moving to the next trial.

Fig. 2. Links

Motion (Figure 3): A small text box with navigation links at the top (home, about, portfolio, studies, contact) is presented, and the subject is instructed to click on one randomly selected navigation link. In normal trials, clicking on any navigation link would move to the next trial, and in error trials, the box would rapidly vibrate for 1.2 seconds before moving on to the next trial.

Fig. 3. Motion

Buttons (Figure 4): A large red button is presented on screen and the subject is instructed to simply click on the button. In normal trials, the button depresses on click and then move on to the next trial. In error trials, the button does not depress, and there is a 1.2 second delay before moving to the next trial.

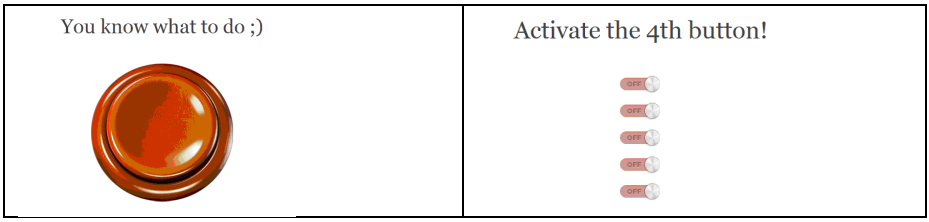


Fig. 4. Buttons

Sound (http://jenius.me/sound.html): A vertical row of 5 sliding switches in the "off" position are presented and the subject is instructed to activate one randomly selected switch. In normal trials, the switch slides to the "on" position and it advances to the next trial. In error trials, the switch slides to the "on" position, then emits a loud siren noise for approximately 2 seconds before moving on to the next trial.

4 Results

4.1 Statistical Analysis

We analyzed the Sternberg memory task data using BATCH software (Advanced Brain Monitoring), splitting the data into ERP windows extending 1.5 seconds before after the subject's response. Subsequent analysis revealed that only the data after the stimulus was necessary, so we cut out the 1.5 seconds prior to the stimulus. We then took a folding average across similar trials ('error potential expected', when the subject had an incorrect answer, and 'no error potential expected', when the subject had a correct answer) and across subjects for all electrodes.

Previous studies have shown that error-related negativity can be reliably detected by a significant difference between the EEG signal on error trials vs. non-error trials along the fronto-central midline, principally the Fz and Cz electrodes (Pailing et. al. 2002). We adopted this method and compared the cross subject and cross trial averages for the Cz and Fz electrodes in terms of error trials and non-error trials, using an ANOVA at 95% confidence to confirm significance. There was a significant difference found between error trial and non-error trial conditions at the $p<.05$ level for the Cz electrode [$F(1, 7) = 4.15$, $p = 0.421$]. We did not find a significant difference at the Fz electrode, but previous studies have shown that a significant difference at Cz is sufficient to confirm the presence of an error potential (Pailing et. al. 2002).

Next we conducted analysis on the data generated by the interface tasks comparing the EEG data when the UI functioned properly versus the EEG data when the UI

was not functioning correctly. Results indicate that we were able to identify error potentials in the Motion, Sound, and Symbols conditions, but not in the Links or Buttons conditions (see Table 1).

Table 1. ANOVA values discriminating error and no-error conditions

Condition	*Results*
Symbols	F(1,7)=40.78, p=2.9e-10
Motion	F(1,7)=210.44, p=0
Sound	F(1,7)=127.62, p=0

4.2 Machine Learning

Our end goal is to be able to use this technology to conduct analysis in real-time, creating the opportunity for integrated methods (e.g., an observation while measuring with the EEG). In order to reach that goal we will need to be able to classify the data using machine learning techniques. Post-hoc analysis will always be useful; to that end we recently were able to synchronize an EEG and eye-tracking device and screen recorder with high temporal precision in our lab. However, we do not consider post-hoc analysis to be a sufficient stopping point.

For this study we implemented a Naïve Bayes Classifier in Matlab, randomly partitioning the Sternberg memory data equally for training and testing. Furthermore, for each 1 second window, we concatenated the Cz and Fz data together for input into the classifier. We repeated this process 10 times, with different partitions of data in our training and testing sets. The results were promising, showing an average of 75.8% correct across all subjects. We did the same machine learning on the data from the UI error and no-error conditions. Results averaged 69.2% across subjects.

5 Discussion

We hypothesized that users encountering interface errors will produce an ERN. This hypothesis was partially supported. Although the symbols, sound, and motion conditions produced significantly different EEG data, the button and link conditions did not. One thing that the button and link conditions had in common is that they were the only UI design flaws that included a *lack* of feedback. The other UI conditions included motion, sound, and symbols, which all included more overtly incorrect feedback. Perhaps the less overt broken links and buttons did not produce a strong enough ErrP for us to identify with our EEG.

Based on these results, we argue that the ErrP would be a valuable addition to the methodological toolbox of usability testing experts focusing on all types of systems from web sites to game design. Although the ErrP does not catch all interface errors, it will catch many of the interaction errors where users tend to blame themselves,

and thus perhaps not report the error. Thus, the ErrP offers a way to systematically and quantitatively review a system for difficult to detect errors that lead to user frustration, dissatisfaction, and reduced performance.

We also note that using the ErrP, and EEG in general, for usability testing is still relatively rare. Much more work can be done to explore the different ways in which other EEG signals can be incorporated into the usability testers' toolbox, particularly as a way to counter the limitations of existing methods that rely on self-report measures.

Other EEG signals, including Mismatch Negativity, N2pc, P300, and P3a/b, Alpha, Beta, and Theta rhythms could comprise a more well-rounded set of measurements for evaluating interface usability directly through brain monitoring, capturing a larger portion of the errors. An achievement such as this would dramatically increase the quality of usability testing, and done in conjunction with other more standard methods, provide substantial feedback leading to significant increases in user experience.

References

Coles, M., Rugg, M.D.: Event-related brain potentials: an introduction. Electrophysiology of Mind. Oxford Scholarship Online Monographs, pp. 1–27 (1996)

Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J.: ERP components on reaction errors and their functional significance: a tutorial. Biological Psychology (2000)

Ferrez, P.W., Millán, J.R.: You Are Wrong! - Automatic Detection of Interaction Errors from Brain Waves. In: Proceedings of IJCAI 2005, pp. 1413–1418 (2005)

Gehring, W.J.: The error-related negativity: Evidence for a neural mechanism for error-related processing. Dissertation Abstracts International 53(10-B), 5090–5090 (1993)

Gehring, W.J., Goss, B., Coles, M.G.H., Meyer, D.E., Donchin, E.: A Neural System for Error Detection and Compensation. Psychological Science 4(6), 385–390 (1993)

Hirshfield, L.M., Bobko, P., Barelka, A., Hirshfield, S., Hincks, S., Gulbrunson, S., Farrington, M., Paverman, D.: Assessing Trust and Suspicion in Human-Computer Interactions Using Non-Invasive Sensors. Tech Report (2013)

Hirshfield, L.M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E.T., Jacob, R.J.K., Sassaroli, A., Fantini, S.: Combining Electroencephalograph and Functional Near Infrared Spectroscopy to Explore Users' Mental Workload. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) FAC 2009. LNCS, vol. 5638, pp. 239–247. Springer, Heidelberg (2009)

Miltner, W.H., Braun, C.H., Coles, M.G.: Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. J. Cognitive Neuroscience 9(6), 788–798 (1997)

Mouraux, G.D., Iannetti: Across-trial averaging of event-related EEG responses and beyond. Magnetic Resonance Imaging 26, 1041–1054 (2008)

Nieuwenhuis, S., Ridderinkhof, R., Blom, J., Band, G.P.H., Kok, A.: Errorrelated brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. Psychophysiology 38, 752–760 (2001)

Oliveira, F.T., McDonald, J.J., Goodman, D.: Performance monitoring in the anterior cingulate is not all error related: Expectancy deviation and the representation of action-outcome associations. Journal of Cognitive Neuroscience 19, 1994–2004 (2007)

Pailing, P.E., Segalowitz, S.J., Dywan, J., Davies, P.L.: Error negativity and reponse control. Psychophysiology 39, 198–206 (2002)
Shneiderman, B., Plaisant, C.: Designing the User Interface: Strategies for Effective Human-Computer Interaction, 4th edn. Addison-Wesley, Reading (2005)
Sternberg, S.: High speed scanning in human memory. Science 153(736), 652–654 (1966)
Tan, D., Nijholt, A. (eds.): Brain-Computer interfaces: Applying our minds to human-computer interaction. Springer, Heidelberg (2010)
Tatum, W.O., Husain, A.M., Benbadis, S.R.: Handbook of EEG. Interpretation. Demos Medical Publishing (2008)

An Effective ERP Model for Brain Computer Interface

Mariko Funada[1], Yoshihide Igarashi[2], Tadashi Funada[3], and Miki Shibukawa[4]

[1] Hakuoh University, Department of Business Administration, Oyama, Tochigi, Japan
[2] Gunam University, Professor Emeritus, Kiryu, Gunma, Japan
[3] Rikkyo University, College of Science, Toshima-ku, Tokyo, Japan
[4] Hakuoh University, Department of Education, Oyama, Tochigi, Japan
mfunada@fc.hakuoh.ac.jp

Abstract. The investigation of BCI (Brain Computer Interface) is particularly interesting for HCI research. Some of recent results concerning BCI have much contributed to the progress of HCI. In this paper we propose an effective ERP model that can reduce the difference among individuals in the process of repetitive tasks. Human brain reactions are quantified by ERPs (Event Related Potentials) that reflect the change of brain reactions through repetitive tasks. We discuss a method of how to even out the difference appeared in ERPs among individuals.

Keywords: BCI, EEG, ERP, Individual difference, Model.

1 Introduction

In order to achieve effective HCI, it is important for us to consider how to treat the difference among individual human beings. Since brains control the behavior of human beings, studies on BCI are indispensable for HCI research. In our experiments, we use monotonous repetitions of tasks by subjects. We recognize some notable changes of ERPs[3,4,5] during the engagement of repetitive tasks by subjects. We propose an effective ERP model that reflects the change of brain reactions through monotonous repetitions of tasks. Using the model we discuss the difference of ERPs among the individual subjects and the possibility of reducing the effect of the difference by taking the average of the ERPs. In this way, we propose a method of how to reduce the effect of the difference among individuals in the process of repetitive tasks.

As monotonously repetitive tasks in our experiments, we choose division questions such that the correct answer to each question can be easily and uniquely determined by subjects. Since EEGs (Electroencephalograms) are relatively easily measured, we adopt them as useful information from the brains in our experiments. In order to investigate the relation between the task effects and EEGs, we use ERPs that are normalized potentials caused by the brain reactions to tasks by subjects.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 299–307, 2013.

2 Methods

2.1 An Experimental Method

We repeat each of the following experiments from five to eight times:

- The subjects are 5 right-handed men and 3 right-handed women from 20 to 23 years old. We identify each of the subjects by *a* to *h*.
- The place of the experiments is the laboratory of the first author at Hakuoh University.
- We use two kinds of stimuli. As shown in Fig. 1, one is a division question (Fig.1 (a)) and the other is a circle (Fig.1. (b)). Each stimulus is displayed in the size of 80×240 pixels.

(a) a stimulus of a division (b) a stimulus for asking to input an answer

Fig. 1. Examples of a division question and a stimulus for asking to answer

- A subject calculates a division when the division question is displayed, and he/she inputs the answer to the question when a circle is displayed. We call the calculation work by a subject "a task".
- Each stimulus of a sequence of stimuli is displayed sequentially in a CRT (Cathode Ray Tube) of 19 inches placed in front of a subject. The number of repetitions of tasks (i.e., the number of division-and-circle stimuli in the sequence) is 100. A subject watches each stimulus without moving his/her eyes. As a stimulus, a division question or a circle for answering (as shown in Fig. 1) is displayed for 1 sec. The time interval between two consecutive stimuli is randomly chosen within the range from 800 [ms] to 1200 [ms].
- The EEGs as a response to a set of division questions and circles for answering are recorded in real time (strictly speaking, it requires 1 sec to record them). Consequently about 6.7 minutes are required for one experiment (i.e., 100 division questions together with answering).
- The single polar and eight channels of the "International 10-20 method" are used for the measurement of EEGs. The positions of the measurement are at C3, C4, Cz, and Pz. The base is A1 that is connected to A2. The sampling frequency for the A/D converter is 1 kHz. We mainly analyze EEGs recorded at Cz.

2.2 An Analytical Method

The recorded EEGs are filtered by an adaptive filter, and the EEGs are normalized by taking the average of their waveforms and by using the standard deviation of the data. Then we measure the ERPs of 100 repetitions of tasks by using the normalized EEGs, the AM (Averaging Method), and the DSAM (Data Selecting and Averaging Method)[1]. Furthermore, we measure the ERPs by using every set of 21 data and then taking their average (Moving Average Method, MAM).

3 Results

3.1 The Ratios of Correct Answers

The ratios of correct answers (RCA) of eight experimental days are depicted in Fig.2. The horizontal axis indicates the experimental days and the vertical axis indicates the RCA. The subjects are divided into four groups: {*c*}, {*b*, *d*, *f*, *g*, *h*}, {*e*} and {*a*} by the Cluster Analysis. The order of groups is the order of RCA.

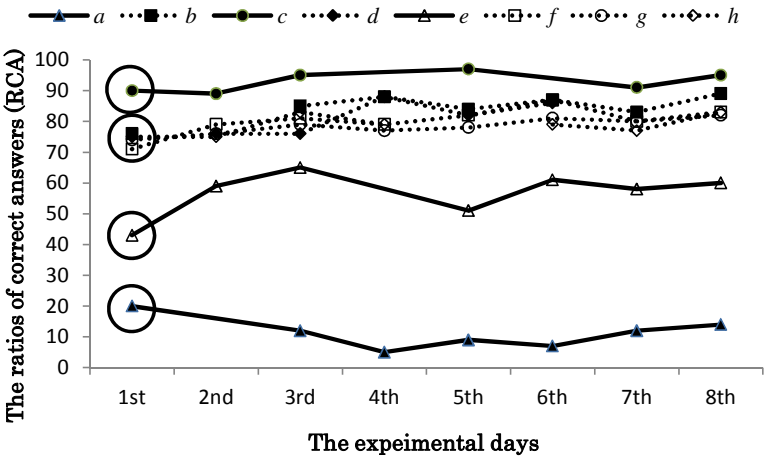


Fig. 2. The ratios of correct answers (Subjects: *a* - *h*)

3.2 ERPs Obtained by the AM and Patterns of ERPs

We calculate ERPs for each experimental day by the AM, and averaging them, we obtained ERPs as shown in Fig.3. The horizontal axis indicates the time after a stimulus is given, and the vertical axis indicates the amplitude of ERPs. The ERPs in Fig.3

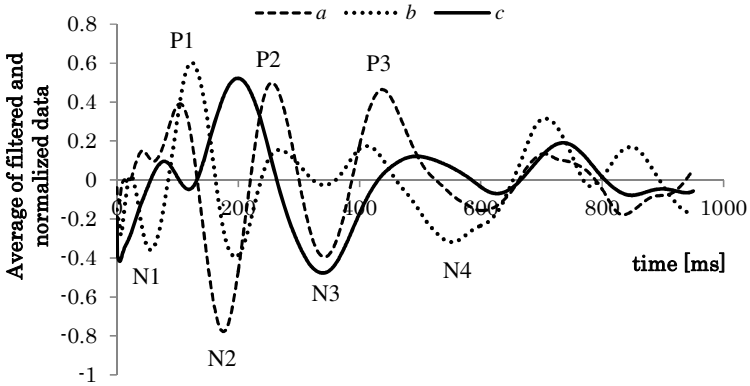


Fig. 3. An example of ERPs (subjects: *a* - *c*)

are from subjects *a*, *b* and *c*. The negative potentials and positive potentials called N1, P1, N2, P2, N3, P3 and N4 appear alternatively. The differences of the waveforms of the ERPs among these subjects are comparatively large. The parts of waveforms N1-P1-N2-P2-N3 are remarkably different, but the latencies of N3 are nearly the same among subjects.

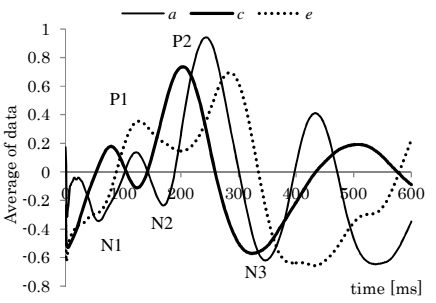
Next we select all EEG*i*(*t*)'s and average them. Then we obtain an ERP, where "*i*" is the question number when the subject finds the correct answer. In Fig.4 (a) and (b) we show ERPs when correct answers are found. The waveforms N1-P1-N2-P2-N3 in Fig.4(a) are resemble each other but the latencies are different. The latencies of the waveforms N1-P1-N2-P2-N3 in Fig.4(b) are almost the same. So we average the ERPs in Fig.4(b) and obtain four patterns of the waveforms N1-P1-N2-P2-N3 as shown in Fig.5. Pattern *A* is obtained from subject *c* in Fig.4(a), pattern *B* is obtained from Fig.4(b), patterns *C* and *D* are obtained from subjects *e* and *a* in Fig.4(a), respectively. These four patterns correspond to the ratios of correct answers (RCA): the RCA of patterns *A*, *B*, *C* and *D* are 92.8%, 80.7%, 56.7% and 11.3% respectively.

We consider that an ERP obtained from a subject for a task corresponds to his/her proficient state (or RCA) of the subject. If the proficient state (or RCA) is improved, then the ERP reflects the progress of the subject. So we consider the following model to represent the changes of ERPs:

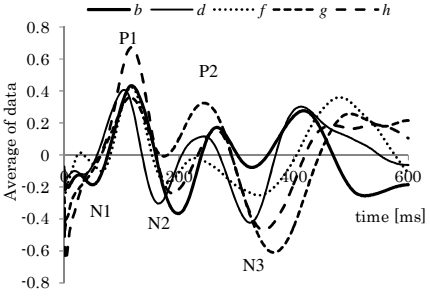
$$ERP(t) = \sum_{k=1}^{5} w_k f_k(t - t_k) \tag{1}$$

where $f_k(t - t_k)$ is the kth potential, w_k is the weight of $f_k(t - t_k)$, and t_k is the latency of the kth potential (k = 1, 2, .., 5). ($t_1 < t_2 < \ldots < t_5$). The term $w_1 f_1(t - t_1)$ represents potential N1, $w_2 f_2(t - t_2)$ represents P1, and so on. We can estimate the component of $w_k f_k(t - t_k)$ by using a normal distribution as follows[2]:

$$w_k f_k(t - t_k) = (-1)^h w_k \frac{1}{\sqrt{2\pi} s_k} \exp\left(-\frac{(t - t_k)^2}{2 {s_k}^2} \right)$$



(a) ERPs when subjects *a*, *c*, and *e* found correct answers



(b) ERPs when subjects *b*, *d*, *f*, *g* and *h* found correct answers

Fig. 4. ERPs in the case of correct answers

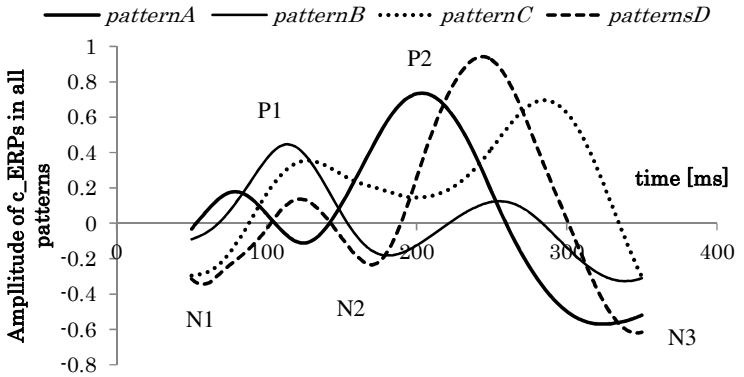


Fig. 5. Four patters of the part N1-P2-N2-P2-N3 in the case of correct answers

3.3 ERPs by the DSAM and Distributions of Potentials

An example of potential distributions used by the DSAM is shown in Fig.6. The bold curve and the dotted curve are the distribution of potentials on the 3rd and 6th experimental days, respectively. The result shows the maximum frequency changes through the repetitions of experiments.

Using the same data in Fig.6, ERPs in Fig.7 are obtained by the DSAM. As the maximum frequency varies, the waveforms of ERPs vary. The distributions and ERPs of all subjects vary through the repetition of experiments.

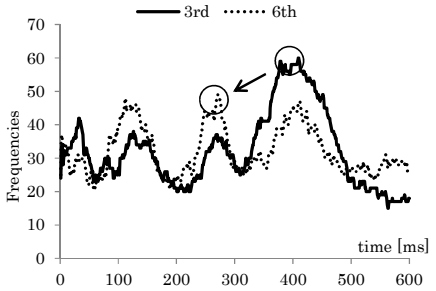


Fig. 6. The potential distributions on the 3rd and 6th experimental days (Subject: *b*, repetition: 100 times)

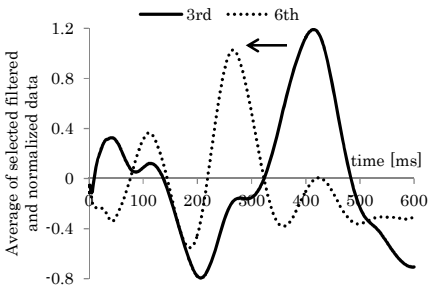


Fig. 7. ERPs obtained by DSAM (subject: *b*, number of data: 59 (3rd), 48 (6th))

3.4 ERPs by the MAM and Their Changes

We calculate ERPs by moving every set of 21 data. Then we obtain ERPs in Fig.8. The ERPs are the 11th, 21th, …, 81th, 90th waveforms. Though N1, P1, N2, P2 and N3 appear on every waveform, the latencies and amplitudes continuously change through the repetitions.

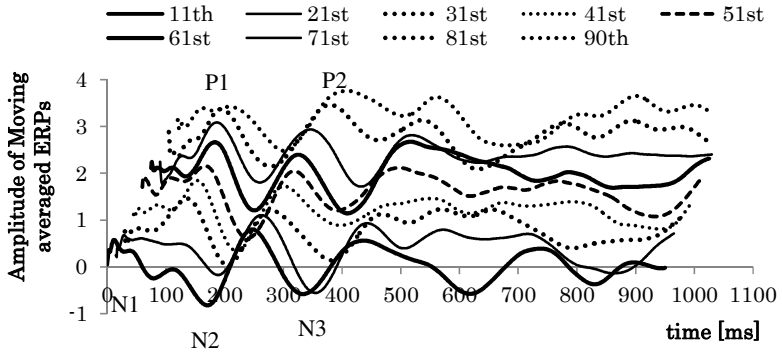


Fig. 8. An example of moving averaged ERPs (subject: *a*, 1st experimental day)

Applying the model (1) to the patterns in Fig.5, we determine the values of all parameters. In Table 1 we show the parameters of the pattern *A*. We calculate cross-correlations between the revised patterns and all moving average of ERPs. In Fig.9 we show the stacked chart of all cross-correlations. The ERPs continuously change through the calculations. In Fig.10 we show the cross-correlations and moving average of RCA. The relation between cross-correlation and RCA suggests some possibility of estimating the values of RCA using the values of cross-correlations.

Table 1. The parameters of pattern A using the model(1)

Potentials	k	h_k	w_k	t_k	s_k
N1	1	1	6.657	30	12
P1	2	0	4.424	82	16
N2	3	1	9.313	129	25
P2	4	0	53.722	201	41
N3	5	1	85.177	340	71

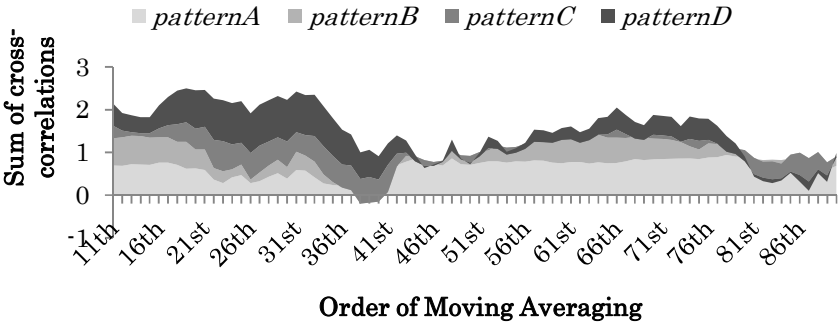


Fig. 9. An example of cross-correlation between moving averaged ERP and four patterns (subject: *a,* 1st experimental day)

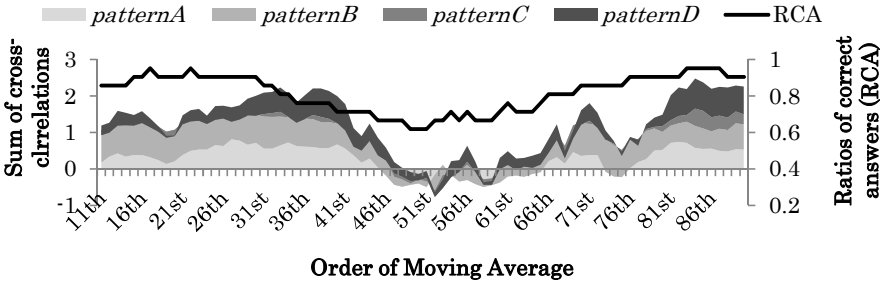


Fig. 10. Another example of cross-correlations between moving average ERP and four patterns, and moving average ratios of correct answers (subject: *d*, 5th day)

3.5 Models for Estimating the RCA

We define 4-tupple (x_A, x_B, x_C, x_D) as the values of cross-correlations between the revised pattern A, B, C and D, respectively. We also define moving averaged ERPs. Defining the moving averaged RCA as a criterion variable, and x_A, x_B, x_C, x_D as explanatory variables, we analyzed the data (RCA, x_A, x_B, x_C, x_D) by Regression Analyses (RA). An example of the results of RA is shown in Fig. 11. The adjusted coefficient of determination is 0.91, and the estimation is good. But the results are different among experimental days, and among the subjects.

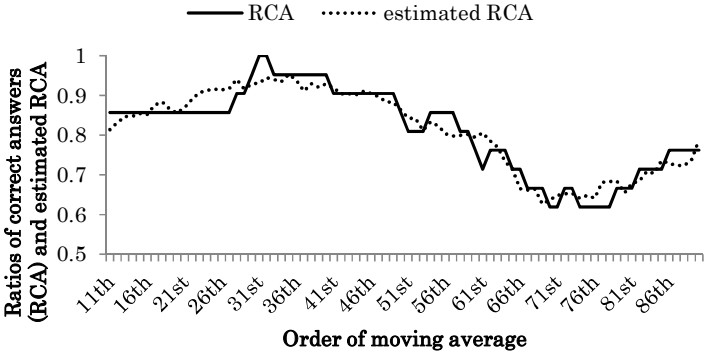


Fig. 11. An example of ratios of correct answers (RCA) to estimated RCA (subject: *b*, 7th experimental day)

We calculate the average of cross-correlations between the revised patterns and the RCA. Concerning the average of pattern *B*, the tendencies in lower RCA (line (1) in Fig. 12) and higher RCA (line (2) in Fig. 12) are very different. We select the higher RCA data, and analyze them by Stepwise Regression through Backward Elimination: the criterion variable is the average of RCA, and the explanatory variables are initially four cross-correlations. The results of the adjusted coefficients of determination (R^2) are shown in Fig. 13. Using the pattern *B* as an explanatory variable, the estimation of RCA is shown in Fig.14.

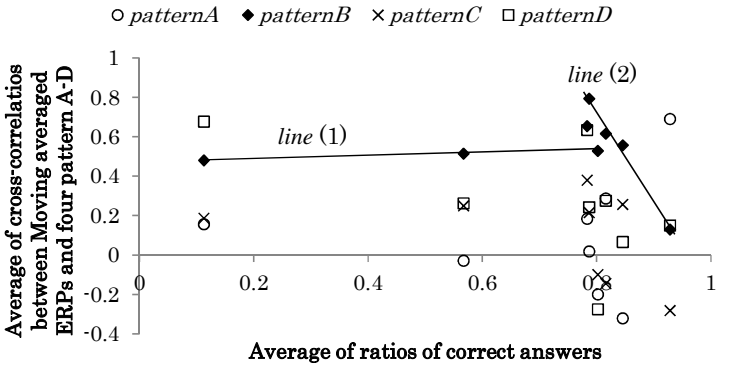


Fig. 12. Averaged RCA and the average of cross-correlations between moving averaged ERPs and four revised patterns

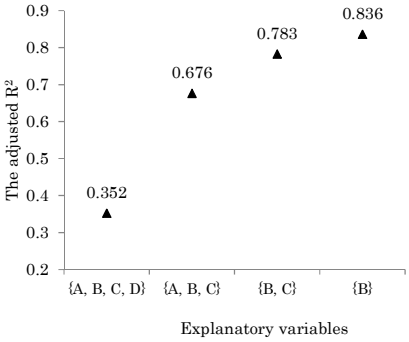


Fig. 13. The adjusted R^2 of Stedpwise Regression

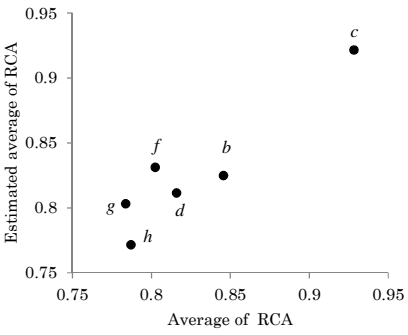


Fig. 14. Result of Stepwise Regression using the pattern *B* as an explanatory variable

4 Discussions

From our results, the individual differences included in ERPs are caused partially by repetition of calculations in each experiment and partially by repetition of experiments. The change of ERPs is mainly caused by the change of distribution of potentials. In the group of proficient subjects for the task (in our experiments, subjects with the RCA greater than 78%), the characteristics or similarity of ERPs have a linear relation. The results in this paper explain that the individual difference appearing in ERPs can be expressed as the following equation:

$$y = f(\text{the level of proficiency for a task}) + g(\text{others}) \quad (2)$$

In our experiments, f (the level of proficiency for the task) is almost 83.6% of y. In other words, the individual differences can be reduced to about 83.6% statistically.

5 Conclusions

From our analysis and discussions we come to the following conclusions:

- We have proposed a method for reducing the individual differences in ERPs.
- The method can be used to clarify various patterns of ERPs for the tasks. We can estimate the similarity of correct answer ratios by cross-correlations.
- The individual differences caused by the proficiency for the task can be reduced by using our method.

Acknowledgement. This research has been partly supported by a Grant-in-Aid for "Challenging Exploratory Research" (No. 23650549) from Japan Society for the Promotion of Science.

References

1. Funada, M., Shibukawa, M., Igarashi, Y., Funada, T., Ninomija, S.P.: Some characterizations of event-related potentials by reconstructing the distribution of the potentials. JapaneseJournal of Human Factors 46(2), 144–156 (2010)
2. Funada, M., Igarashi, Y., Funada, T., Shibukawa, M.: A Model Reflecting the Changes of Erps During Repeated Learning of Calculations. In: The Second IASTED Asian Conference onModelling, Identification and Control, Phuket, Thailand, pp. 769–86 (2012)
3. Funada, M., Funad, T., Shibukawa, M., Akahori, K.: Quantification and Analysis of Efficiency of Iterative Learning by Using Event Related Potentials. Educ. Technol Res. 35, 103–113 (2012)
4. Luck, S.J.: An Introduction of the Event-Related Potential Technique. The MIT Press, Cambridge (2005)
5. Picton, T.W., Bentin, S., Berg, P., Conchin, E., Hillyard, S.A., Johnson Jr., R., Miller, G.A., Ritter, W., Ruchkin, D.S., Rugg, M.D., Taylor, M.J.: Guideline for using human event-related potentials to study cognition: recording standards and publication criteria. Psychophysiology 37, 128–152 (2000)

Neural Oscillatory Signature of Original Problem Solving

Henk J. Haarmann[1,*], Polly O'Rourke[1], Timothy George[1], Alexei Smaliy[2], Kristin Grunewald[1], and Joseph Dien[1]

[1] Center for the Advance Study of Language, University of Maryland, College Park, MD, USA
{hhaarmann,porourke,tgeorge,kgrunewald,jdien}@casl.umd.edu
[2] Maryland Neuroimaging Center, University of Maryland, College Park, MD, USA
asmaliy@umd.edu

Abstract. The goal of the present research was to increase understanding of the neural oscillatory signature of originality in verbal divergent thinking by determining if event-related synchronization (ERS) in frequency bands other than alpha predicts originality. EEG was recorded while participants performed the insight task in which they were presented with a brief scenario and asked to generate as many explanations as possible during a three minute period. After the EEG session, participants were asked to rate the originality of each idea they produced. Analyses revealed that high originality was associated with decreases in the high beta ERS and with hemispheric asymmetry in the low beta band, immediately prior to idea generation. These results suggest the neural signature of originality extends beyond hemispheric asymmetries in the alpha band and provide important insights into the neural underpinnings of verbal creativity.

Keywords: Divergent thinking, originality, EEG, ERS, alpha, beta.

1 Introduction

Divergent thinking is a type of creative problem solving which involves the generation of multiple, distinct solutions to open-ended problems. These solutions will vary in their level of originality such that some will be highly unique, unusual ideas while others will reflect more standard approaches. One commonly used task is the insight task (IS) in which participants are presented with a brief scenario ("a light in the darkness") and asked to provide as many explanations as possible within a specified time period. A standard, fairly unoriginal response would be "headlights on a car", while "jelly fish in the ocean" is an example of a highly original solution. There is a substantial literature examining the neural oscillatory signature of divergent thinking [1] which suggests that original divergent thinking is associated with increases in spectral power in the alpha band (8-12 Hz). Alpha power is inversely related to neural activation such that greater alpha power indicates reduced activation and vice versa.

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 308–315, 2013.

One standard view of alpha is that it reflects cortical idling [2]. It is believed that this disengagement is what enables the formation of atypical conceptual combinations between weakly related or unrelated concepts. Increases alpha power have also been found during cognitive activity as a function of task demand [3-4]. The dominant theory is that alpha reflects a top-down process that inhibits the intrusion of sensory information or conflicting operations [3], [5-7]. In the case of verbal creativity tasks, such as the remote associates test (RAT), the increased alpha power is thought to inhibit the activation of words and concepts that are strongly associated with the cue thus reducing the difficulty of selecting weak associates in the face of competition [8].

One focus of research on the neural basis of divergent thinking is to determine what distinguishes highly original and less original ideas at the neural level. A recent study by Grabner, et al. [5] found that (self-rated) originality of ideas generated during divergent thinking is predicted by event-related synchrony (ERS) in the alpha band over the right hemisphere during idea generation. In their study, highly original ideas were associated with increased alpha ERS in the right hemisphere relative to less original ideas. Their study did not, however, report analyses of whether power in other frequency bands, such as the beta band, predicts originality. The goal of the present study was to fill that gap by determining if other frequency bands are sensitive to originality level during idea generation, and examine how any effects interact with hemisphere given that, though the literature is not unambiguous, there is a strong suggestion of greater engagement of the right hemisphere during the generation of highly original ideas [5], [9].

2 Methods

The current analysis includes data from two experiments, both of which employed the IS as the divergent thinking task.

2.1 Participants

Data from a total of 41 neurologically normal participants were included in this analysis. 21 (7 male; mean age 21.1, S.D. 2.3) from Experiment 1 and 20 (10 male; mean age 21.0, S.D. 1.5) from Experiment 2. All participants were right handed native speakers of English.

2.2 Materials

In both experiments, participants performed the insight (IS) task, in which they were presented with situations and asked to produce different explanations. The following are the test items:

- “a light in the darkness”
- “Person A is lying down, person B is sitting and person C is standing”
- “a cloth in the air”
- “Person A walks, Person B jumps”

Items 1 and 2 are the English translations of the items used in [5]. Items 3 and 4 were added in order to increase the power of the design.

2.3 Procedure

After signing a consent form and electrode application, participants were tested individually in a sound-attenuating room with the lights turned off. Participants were seated in a comfortable chair in front of a computer monitor and asked to use a chin rest in order to minimize movement artifacts. Prior to the task, two one-minute base baselines were recorded, one with eyes open and the other with eyes closed. Each of the four task items consisted of the presentation of a fixation cross for 15 seconds, followed by the presentation of the item. Participants were instructed to generate as many solutions as possible and to be creative in their responses. When participants had an idea, they pressed a button on the response box. They then vocalized their idea and pressed the button again to indicate when they were done. The response period for each item lasted 3 minutes. The testing session lasted approximately 30 minutes. At the conclusion of the EEG session, participants were presented with the transcription of the ideas they produced for each item and were asked to rate the originality of each response on a scale of 1 to 5. In Experiment 1, participants were exposed to pink noise during task performance while in Experiment 2 there were no auditory stimuli. Also, in Experiment 2, participants had a one minute rest between the IS task items.

2.4 EEG Recording

Electroencephalographic (EEG) data were acquired with a 128-channel HydroCel Geodesic Sensor Net using the Electrical Geodesics Inc. (EGI) NetStation system. The EEG signal was sampled at 250 Hz. The signal was high-pass filtered online at 0.1 Hz, low-pass filtered at 100 Hz, and notch filtered at 60 Hz. Impedances were kept below 50 KΩ where possible per manufacturer recommendation, and otherwise under 100 KΩ.

2.5 EEG Analysis

EEG data were artifact-corrected using the EP toolkit for MATLAB [10]. Spectral power was obtained through Fast Fourier Transform averaged across 1-second epochs within a period. By-subject averages of EEG spectral power were obtained for every cell in our design and then log-transformed, except for power values that were entered into the ERS analyses [cf. 5]. ERS was calculated using the following formula [11]: %ERS = [(Activation-Reference)/Reference] x 100. The 1000 ms period terminating 250 ms prior to first button press (indicating an idea) served as the activation interval. Reference was the pre-task eyes-open baseline.

The design for the data analysis was the following: Originality (High, Low) x Hemisphere (Left, Right) x Lobe (frontal, temporal, parietal, occipital). This repeated measures ANOVA was performed for each frequency band with Experiment as a

between subjects factor. The lobe by hemisphere division of the scalp electrodes was accomplished via Brain Voyager QX 2.4 (Brain Innovation, Maastricht, The Netherlands) which mapped the 10/20 sensor positions, using the coordinates set forth in [12] on to brain lobes. The frequency bands were defined as follows: delta (1-4 Hz), theta (4-8 Hz), lower alpha (8-10 Hz), upper alpha (10-12 Hz), overall alpha (8-12 Hz), low beta (12-16 Hz), mid beta (16-20), high beta (20-28 Hz) and gamma (28-70 Hz). The EEG data for each response in the Insight Task was categorized as high or low originality (as per the median split performed on the originality ratings). The degrees of freedom in all analyses were Greenhouse-Geisser corrected when appropriate.

In addition, a correlational analysis, using Pearson product-moment correlations, was run using an index of hemispheric asymmetry (right minus left hemisphere) for the alpha and beta sub-bands in order to assess the impact of hemispheric asymmetry on divergent thinking performance (fluency and originality). Fluency was defined as the number of distinct ideas and originality as a rating from 1 (least original) to 5 (most original). The analysis was run on both the pooled data from the two experiments and on each experiment separately.

3 Results

When analyzed separately, the data from the two experiments showed effects of originality in the alpha and beta bands only. These two bands will, therefore, be the focus of this analysis. In the lower alpha band, there was a four-way interaction of Originality, Lobe, Hemisphere and Experiment ($F(3,117) = 3.24$, $p < .05$, $\eta_p^2 = .077$). Splitting across the factor Experiment revealed a significant Originality x Hem x Lobe interaction ($F(3,57) = 2.935$, $p = .050$, $\eta_p^2 = .134$) in Experiment 2 only. Separate analysis of the levels of the Lobe factor reveals an interaction of Originality and Hemisphere in the frontal lobe ($F(1,19) = 7.431$, $p < .05$, $\eta_p^2 = .281$) such that high originality responses had greater ERS in the left hemisphere while ERS for low originality responses was greater in the right. Simple comparisons revealed no significant effects. In the high alpha band, there was also a significant four-way interaction ($F(3,117) = 4.14$, $p < .05$, $\eta_p^2 = .096$). This was driven by a marginally significant Originality x Hem x Lobe interaction ($F(3,60) = 3.183$, $p = .059$, $\eta_p^2 = .137$) in Experiment 1 only. Splitting across the factor lobe revealed a marginal interaction of originality and hemisphere in frontal sites ($F(1,20) = 4.038$, $p = .058$, $\eta_p^2 = .168$). Simple comparisons showed a main effect of hemisphere for high originality responses in the frontal lobe ($F(1,20) = 4.899$, $p = .039$, $\eta_p^2 = .197$) such that ERS was greater in the right hemisphere than the left. Finally, in the high beta band there was a main effect of Originality in the high beta band ($F(1,39) = .7863$, $p = .008$, $\eta_p^2 = .185$), such that ERS was greater for low originality responses pooled across both experiments. When the data from each experiment were analyzed separately, the effects of originality on the high beta band were marginal (Experiment 1, $F(1,20) = 4.129$, $p = .058$, $\eta_p^2 = .171$; Experiment 2, $F(1,19) = 4.320$, $p = .051$, $\eta_p^2 = .185$) but in the same direction as the pooled analysis.

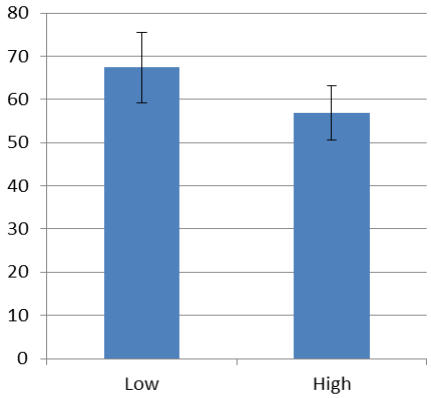


Fig. 1. Percentage of ERS for low and high originality responses in the high beta band with standard error bars

The correlational analysis ran between the index of hemispheric asymmetry and the behavioral measures. In Experiment 1, there was a positive correlation between hemispheric asymmetry and fluency in the high alpha band ($r = .545$, $p < .05$). In Experiment 2, there was a positive correlation between asymmetry in the low beta band and originality ($r = .488$, $p < .05$). The pooled analysis did not yield any significant correlations.

4 Discussion

The results of the current experiment provide a more nuanced view of the neural signature of originality. While confirming the association between alpha power and originality (albeit only significant in Experiment 1), the finding that the beta band is also sensitive to originality level is a novel finding. Similarly, the suggestion that hemispheric asymmetry in not only the alpha but also the beta band during idea generation predicts performance underlines the contribution of beta to divergent thinking.

The findings for the ERS in the alpha band in the two experiments are somewhat contradictory in terms of sub-band and hemispheric distribution. In Experiment 1 the ANOVA results showed that highly original responses had increased high alpha ERS in the right compared to left frontal areas. In experiment 2, a marginal interaction of originality and hemisphere in frontal areas for the low alpha band suggested that highly original responses elicited increased ERS in the left hemisphere. It is possible that the presence or absence of pink noise contributed to this difference but not likely given that the pattern exhibited in Experiment 1 is very similar to that found by Grabner et al. [5] in an experiment that contained no auditory stimuli. The discrepancy is, however, not surprising as in the wider literature findings for the topographical distribution of alpha tend to be inconsistent [1]. It must be recognized that both experiments did show evidence of an association between originality and frontal alpha ERS which generally does replicate previous experiments using the IS task [9], [13].

The finding of an association between decreased spectral power in the high beta band and generation of high originality responses was more robust and has not been previously reported. Dietrich and Kanso indicate that "the single most common finding in this literature is the absence of significant changes to the beta frequency" [1 p. 825]. They are also surprising as oscillations in the beta frequencies are associated with motor processes. The execution of movements is associated with desynchrony in the beta band [2], as is observing movements [14] or imagining them [15]. While the production of both high and low originality ideas involve movements and, therefore motoric processes, high originality responses involve retrieving and producing words that are not strongly primed by the cue and may, as a result, involve more effortful production and, therefore, distinct (and perhaps increased) motoric demands, evidenced by increased beta desynchrony. However, while lexical-semantic priming affects the time course of lexical access, it does not necessarily affect subsequent speech motor planning and execution.

The notion that distinct linguistic demands are reflected in power changes in the beta band is supported by recent research in semantic processing [16]. Luo, et al. [17] and Wang, et al. [18] have found decreases in power in the mid beta band (16-19 Hz in [17]. and 16-20 in [18]) for semantically incongruent words in a sentence compared to congruent words. These studies examined brain activity during language perception while we looked at activity during idea generation (preceding speech production) so it is difficult to make direct comparisons with confidence. Nevertheless, as the generation/production of highly original ideas involves accessing and integrating lexical items that are normally not considered related, there is a parallel. [16] and [17] found that unprimed words (semantically incongruent) elicited reductions in beta synchrony, as did the highly original (i.e. unprimed) ideas in our experiments, albeit in the high but not mid beta band. The sensitivity of the beta band to the semantic features of words found in the sentence processing studies listed above implies that our effect may reflect the differing linguistic and cognitive demands of high and low originality ideas.

How specific properties of these demands reflect different sub-bands within the overall beta band remains to be determined. One possibility is that the decreased power associated with high originality in the high beta band reflects decreased demands in active, controlled semantic processing needed for the generation of more original ideas. For example, reducing this type of processing may facilitate access to implicit semantic memory without top-down bias from explicit semantic memory, thus resulting in more original ideas. This explanation is admittedly post-hoc and it does not rule out the possibility that there are conditions under which an increase in active, controlled semantic processing results in greater originality. The latter type of processing can help to resolve the competition between dominant and weak associations in favor of weak ones, thereby contributing to greater originality. This analysis, while speculative, implies that divergent thinking can be achieved with different modes of thinking, which need to be experimentally controlled.

The correlational analysis revealed that in Experiment 1, hemispheric asymmetry (increased ERS on the right compared to the left) in the high alpha band predicted fluency (but not originality), but in Experiment 2 asymmetry in the low beta band

predicted originality (and not fluency). Though the correlations differed across experiments, they do emphasize the importance of hemispheric asymmetry during divergent thinking and provide evidence that fluency and originality are underpinned by distinct neural mechanisms. The relationship between alpha and fluency found in Experiment 1 is novel but generally consistent with Jung-Beeman et al.'s account [8] such that individuals with greater right hemispheric alpha are able to inhibit standard associations in favor of weakly associated concepts while those without the greater right hemispheric alpha may become fixated on standard, unoriginal associations and generate fewer overall responses. The lack of correlation between alpha asymmetry and originality is somewhat surprising. But, as mentioned above, the mode of thinking induced by the experimental context may influence whether increased alpha asymmetry is associated with increased originality. With respect to beta, the correlational analysis of Experiment 2 provides additional evidence for the relationship between beta and generating original responses and suggests that a greater increase in low beta power in the right than left hemisphere may be a key component.

Taken together, these results suggest the neural signature of originality extends beyond the alpha band. The findings that both activity in the high beta band and hemispheric asymmetry in the low beta band predict originality provide new and important insights into the neural underpinnings of verbal creativity. Future research will further elucidate the role of beta in the generation of original ideas. Of particular interest is the question of whether original divergent thinking can arise from different modes of thinking and their associated neurophysiological mechanisms. In this paper, the focus was on the association between neural oscillatory activity and originality. Frequency-specific experimental manipulation of this activity will be crucial for moving beyond association and establishing its causal role in cognitive creativity.

References

1. Dietrich, A., Kanso, R.: A review of EEG, ERP and Neuroimaging Studies of Creativity and Insight. Psyc. Bul. 136(5), 822–848 (2010)
2. Pfurtscheller, G., Stancák, A., Neuper, C.: ost-movement beta desynchronization. A correlated of idling motor area? Electroencephalography and Clinical Neurophys 98(4), 281–293 (1996)
3. Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Res. Rev. 29(2-3), 169–195 (1999)
4. Jensen, O., Gelfand, J., Kounios, J., Lisman, J.E.: Oscillations in the alpha band (9-12 Hz) increase with memory load during retention in a short-term memory task. Cerebral Cortex 12(8), 877–882 (2002)
5. Grabner, R.H., Fink, A., Neubauer, A.C.: Brain Correlates of Self-Rated Originality of Ideas: Evidence From Event-Related Power and Phase-Locking Changes in the EEG. Behav. Neurosci. 121(1), 224–230 (2007)
6. Fink, A., Grabner, R.H., Benedek, M., Neubauer, A.C.: Divergent thinking training is related to frontal electroencephalogram alpha synchronization. Eur. J. of Neurosci. 23, 2241–2246 (2006)

7. Cooper, N.R., Croft, R.J., Dominey, S.J.J., Burgess, A.P., Gruzeiler, J.H.: Paradox lost? Exploring the role of alpha oscillations during externally vs. internally directed attention and the implications for idling and inhibition hypotheses. Int. J. of Psychophys 47(1), 65–74 (2003)
8. Jung-Beeman, M., Bowden, E.M., Haberman, J., Frymiare, J.L., Arambel-Liu, S., Greenblatt, R.: Neural activity when people solve verbal problems with insight. PLoS Biol. 2(4), 500–510 (2004)
9. Fink, A., Grabner, R.H., Benedek, M., Neubauer, A.C.: The creative brain: Investigation of brain activity during creative problem solving by means of EEG and fMRI. Hum. Brain Mapping 30, 734–748 (2009)
10. Dien, J.: The ERP PCA Toolkit: An Open Source Program For Advanced Statistical Analysis of Event Related Potential Data. J. of Neurosci. Methods 187(1), 138–145 (2010)
11. Pfurtscheller, G., Lopes da Silva, F.H.: Event-related EEG/MEG synchronization and desynchronization: basic principles. Clinical Neurophys 110(11), 1842–1857 (1999)
12. Okamoto, M., Dan, H., Sakamoto, K., Takeo, K., Shimizu, K., Kohno, S., Oda, I., Isobe, S., Suzuki, T., Kohyama, K., Dan, I.: Three-dimensional probabalistic anatomical cranio-cerebral correlation via the international 10-20 system oriented for transcranial functional brain mapping. NeuroImage 21, 99–111 (2004)
13. Fink, A., Neubauer, A.C.: EEG alpha oscillations during the performance of verbal creativity tasks: Differential effects of sex and verbal intelligence. Int. J. of Psychophys 62(1), 46–53 (2006)
14. Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., Rizzolatti, G.: Activation of human primary motor cortex during action observation: a neuromagnetic study. Proc. of the Nat. Acad. of Sci. 95, 15061–15065 (1998)
15. Neuper, C., Wortz, M., Pfurtscheller, G.: ERD/ERS patterns reflecting sensorimotor activation and deactivation. Prog. Brain Res 159, 211–222 (2006)
16. Weiss, S., Mueller, H.M.: "Too many betas do not spoil the broth": the role of beta brain oscillations in language processing. Fron. in Psyc. 3, 1–15 (2012)
17. Luo, Y., Zhang, Y., Feng, X., Zhou, X.: Electroencephalogram oscillations differentiate semantic and prosodic processes during sentence reading. Neurosci. 169, 654–664 (2010)
18. Wang, L., Jensen, O., van den Brink, D., Weder, N., Schoffelen, J.M., Magyari, L., Hagoort, P., Bastiaansen, M.: Beta oscillations related to the N400m during language comprehension. Hum. Brain Mapping 33(12), 2898–2912 (2012)

A Real-World Neuroimaging System to Evaluate Stress

Bret Kellihan[1], Tracy Jill Doty[2] , W. David Hairston[2], Jonroy Canady[1], Keith W. Whitaker[2], Chin-Teng Lin[3], Tzyy-Ping Jung[4], and Kaleb McDowell[2]

[1] Intelligent Systems Department, DCS Corporation, Alexandria, VA 22310 USA
[2] Human Research and Engineering Directorate, Army Research Laboratory, Aberdeen Proving Ground, MD 21005 USA
[3] Department of Electrical Engineering and the Brain Research Center, National Chiao Tung University, Hsinchu, Taiwan
[4] Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California at San Diego, La Jolla, CA 92093 USA
{bkellihan,jcanady}@dcscorp.com,
{tracy.j.doty2.ctr,william.d.hairston4.civ,
keith.w.whitaker1.civ}@mail.mil, ctlin@mail.nctu.edu.tw,
jung@sccn.ucsd.edu, kgm8@cornell.edu

Abstract. While the laboratory setting offers researchers a great deal of experimental control, this environment also limits how generalizable the results are to the real world. This is particularly true when studying the multifaceted phenomenon of stress, which often relies on personal experience, a dimension that is difficult to reproduce in the laboratory setting. This paper describes a novel, multi-aspect real-world integrated neuroimaging system (MARIN) optimized to study physiological phenomena in the real-world and particularly suited to the study of stress. This system integrates neurological data from a gel-free, wireless EEG device with physiological data from wireless cardiac and skin conductance sensors, as well as self-reports of activity and stress. Coordination of the system is managed through an Android handheld mobile device that also logs salient events and presents inventories for subjective reports of stress. The integration of these components creates a rich, multimodal dataset with minimal interference to the user's daily life, and these data will guide the further understanding of neurological mechanisms of stress.

Keywords: wireless electroencephalography, skin conductance response, electrodermal activation, heart-rate variability, wearability.

1 Introduction

Understanding the human brain is crucial to the development of technology that will enhance daily life and performance on critical tasks. Our current level of understanding has been vastly improved through increasingly complex and sophisticated laboratory-based experimental research. While this setting grants the researcher a great deal of control, it also limits the ecological validity of the results. For example, it may be difficult to accurately represent phenomena such as fatigue [1], aggression [2], and

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 316–325, 2013.

social preferences [3] within the artificial setting of a laboratory. Therefore, there exists a need to study the brain in its natural environment in order to truly further our understanding of how behavior connects to brain function.

This is particularly true for broad phenomena, such as stress, which involves a confluence of multiple physiological systems that are highly influenced by diverse environmental factors. Stress is defined as anything that disrupts or is perceived to disrupt the complex dynamic homeostasis of the body and brain [4, 5], although we recognize that this may not be a complete definition [6]. In daily life, we face many personal sources of stress that can be difficult to replicate in the laboratory setting, such as stress from one's work environment or stress related to one's family life. Indeed, evidence suggests that physiological responses to stress are larger in the real-world compared to in the laboratory [7, 8]. Meanwhile, increased levels of psychological stress are associated with increased incidence of disease [9] and mortality [10, 11]. Even acute stress can be maladaptive in some individuals [5] and has been shown to affect cognition, although these observations have been limited to laboratory settings [12, 13]. These facts further the importance of real-world research on the effect of stress on the human brain.

This paper discusses the current state of technology available for studying neurophysiological constructs, such as stress, in true "real-world" settings, as well as potential roadblocks that must be addressed in the design of such systems. Here, we discuss efforts focused on developing a real-world neuroimaging system optimized for studying broad-based scientific and applied pursuits of monitoring physiological states, using the study of stress as an exemplar target state. The ultimate goal of this system is to use contextual and physiological information to interpret neurological data. The system described here (MARIN) combines newly developed wireless neuroimaging technology with existing physiological sensors and a mobile user-interface device to record events and collect subjective measurements in real-world environments. MARIN also integrates contextual information from the environment with the high-quality neurological, physiological, and subjective monitoring data. We highlight the specific technological advancements of this device and how it is particularly suited for the study of real-world stress.

2 Background

The physiological response to stress is essential to healthy functioning. This response is considered maladaptive only when it occurs too frequently, is disproportional to the stressor (i.e. chronic stress), occurs in the absence of a stressor, or does not occur when a stressor is present [6]. While a majority of research studies have focused on the deleterious effects of stress, without real-world neuroimaging research, the basic effect of stress on the brains of normal individuals is still unknown. By building upon real-world research of the physiological mechanisms underlying the response to stress, research can begin to make the connection to the natural neurological response to stress.

2.1 Studying Real-World Stress

Although technological advances have only recently enabled real-world neuroimaging of stress in humans, stress has been measured outside of the laboratory via ambulatory cardiac monitoring for some time [8]. Cardiac measurements allow the researcher to tap into the functioning of the autonomic nervous system via measurements of biomarkers, such as heart rate and blood pressure. Stressful stimuli act directly on the autonomic nervous system, generally by activation of the sympathetic nervous system, which mobilizes the body to respond to stress via peripheral physiological functions, such as increased heart rate and sweating [5].

The measurement of skin conductance on eccrine sites (i.e. hands and feet) is another method for assessing autonomic nervous system function, and, therefore, the effect of stress [14]. However, while ambulatory monitoring of cardiac responses in the real-world has been ongoing for several years, the measurement of skin conductance outside of constrained settings has only just begun due to a major challenge to wearability. Traditional skin conductance sensors were placed exclusively on the tips of the fingers, preventing participants from engaging in tasks requiring manipulation of objects. Semi-real-world studies of skin conductance to date have relied upon the subject singularly engaging in a task that does not require dexterity of the fingertips or direct pressure on the soles of the feet, such as driving [15]. However, recent technology has been developed to reliably measure skin conductance via a site on the wrist, allowing the participant a full range of motion [16] and opening the door for integration with a wide range of tasks.

Both cardiac and skin conductance measurements excel at detecting the broad reaction of the body to a stressful event. While this broad categorization of stress is particularly helpful to detect the occurrence of a stressful event, these measures lack a high degree of selectivity, and in fact, are sensitive to many different types of events. Since our goal is to understand how the brain responds to stress, a portable neuroimaging device is crucial to tease these factors apart. Recent advances in neurotechnology have created truly wearable wireless electroencephalography (EEG) systems for real-world research [17] that could be utilized for a host of applications, including stress research. Additionally, a system integrating EEG, heart rate, and skin conductance has been proposed, however, this system has not been designed for real-world experiences; i.e., it is not completely wireless and only features a limited number of electrodes for EEG recording (< 6 channels) [18]. Although recent scientific efforts have been put forth to create stress prediction indices from EEG data [18–21], due to technological limitations these schemes have consisted of laboratory-derived scenarios and have not been utilized in a real-world neuroimaging environment. However, this type of predictive technology would be highly advantageous as part of a wearable EEG system.

2.2 Obstacles for a Real-World Neuroimaging System

We believe one of the most critical components of a real-world neuroimaging system is to create a rich multi-dimensional characterization of context. This allows for the

accurate and meaningful interpretation of measured neural activity. This obstacle has been mitigated in some real-world cardiac monitoring studies by the use of electronic diaries (e.g. [22]).

Additional obstacles fall under three broad topic areas that directly influence the design of such a system, including general wearability, usability for trained (non-scientist) users, and usability for scientific purposes. In order for any system to be suitable for real-life settings, the user must be able to wear it without any substantial hindrance to normal activities. The device must be comfortable enough for the user to wear for multiple hours a day. This means the device should not be too heavy or made of inflexible material. This has been particularly difficult for adapting current EEG acquisition systems for real-world data collection. For example, the system in Figure 1 takes approximately one hour to setup and can be come uncomfortable in minutes. As with any real-world device, usability for the wearer is crucial. The device must be easy for even a trained non-scientist to set up, troubleshoot rare issues with data acquisition, and log events throughout the day. Finally, usability is also important for the scientists analyzing the data. Dropped data packets, time lags, and movement of the different sensors must be minimized. Meanwhile, perhaps most critical for scientific pursuits, raw data from all sensors must be accessible in a manner that facilitates integrated analyses while providing the ability to properly characterize external influences. The system described in this paper has been designed to address these obstacles and create a rich dataset that captures neurological functioning during multiple types of events and states that occur naturally in the real world.

Fig. 1. State-of-the-art laboratory grade EEG system made mobile. The participant wears a high-density wet electrode cap which is wired to a laptop and amplifiers that are placed in a backpack.

3 The Real-World Neuroimaging System

The system developed here (referred to here as "MARIN" – Multi-Aspect Real-world Integrated Neuroimaging system) is a laboratory-grade measurement capability specifically designed to overcome several of the obstacles for real-world neuroimaging, particularly context monitoring (Figure 2). The prototype comprises an Android device, a Samsung Galaxy S III in the current implementation, and three physiological monitoring sensors: a high-density MINDO EEG system (Hsinchu, Taiwan), a multifunction Zephyr Bioharness 3 lightweight chest strap (Waltham, MA, USA), and a multifunction Affectiva Q Sensor wrist-watch style device (Annapolis, MD, USA). The Android device monitors, records, and synchronizes data streamed from the three physiological monitoring devices, as well as obtaining additional user inputs. All three of the physiological systems use dry-type electrodes for quick, easy set-up and longer-term wearability. The complete system weighs approximately 406 grams (0.9 pounds; MINDO-64: 200g, Bioharness 3: 50g, QSensor: 22.7g, Samsung Galaxy S III: 133g).

As this system is designed for scientific pursuits, the primary analysis software will be offline, where the Android-based physiological data and behavioral data can be combined with contextual information from additional sources, such as the user's calendar, user annotations, or questionnaire responses (see below for more detailed description). State-of-the-art offline-analyses programs, such as EEGLAB, will be used for data processing.

3.1 Components

Wireless EEG Cap. The centerpiece of the MARIN System is the NCTU-developed 64-channel wireless EEG system (MINDO-64), which is designed to address high-resolution laboratory-grade data acquisition, long-term comfortable wear, quick user set-up, and high portability. The typical wet electrodes found in laboratory equipment can dry out within 30-minutes to 2 hours, which directly influences signal quality [23]. High-bandwidth data transmission requirements typically force participants to be tethered to computing systems or to carry relatively heavy hardware, such as batteries, amplifiers, and laptop computers [24]. This is especially confounded by the large number of channels typically required (64+) for laboratory-grade research, which may include source localization or separation procedures [25]. These hardware constraints limit the naturalistic behaviors that can be observed, as well as the types of contexts that may be investigated. The MINDO-64 is the first wireless EEG system integrated into a form factor with a flexible printed circuit board inside, a novel head-circumference-adaptable mechanical design for improved stability, and active dry sensors that amplify signals at a very early stage to improve signal-to-noise ratios and avoid the need for skin preparation and gel application. It uses both Bluetooth and WiFi modules to transmit EEG signal during recording, offering a maximum 512Hz sampling rate with 24-bit resolution. Through the integration of active sensor and power control on the main circuit, the system allows long-term wear of up to 10 consecutive hours of operation time. The system's wireless technologies, light weight (<200g), and dry sensor design also support comfort, fast set-up, and portability.

Fig. 2. The MARIN System

Peripheral Monitoring Devices. A Zephyr Bioharness 3 is the system's principal central-physiology sensor suite in MARIN. It is a small, lightweight, self-contained electrocardiography (ECG) system that also provides respiration rate, three-axis accelerometery on the chest, posture, skin temperature and derived measures of heart rate, heart rate RR, breathing rate, and breathing rate RR. The Bioharness is capable of recording all measures except ECG directly on the device; ECG data is transmitted

and stored on an external device (see next section). Together these measures capture several aspects of autonomic nervous system function that can be affected by stress, while the sensor suite can be easily donned and removed by the subject in under a minute, as the sensors are integrated into a single strap that is worn around the chest.

Additionally, MARIN includes the Affectiva Q Sensor, which is a small sensor worn on the wrist, that provides measures of skin conductance via electrodermal activity (EDA) and three axis accelerometery. It is similar in size to a men's wrist watch and attaches with a simple wrist strap. The Q Sensor stores data locally on the device and, similar to the Bioharness, transmits the data to the Android device wirelessly using Bluetooth. While EDA provides an additional modality for autonomic response, the accelerometers can be used to correlate this with modulation in general activity levels (i.e., used as an actigraph).

Handheld Computing Device. An application of constant, day-long monitoring places a premium on lightweight, small form factor systems since the user must carry the computing device with them for the duration of the data acquisition period. For this reason, we chose a cell phone as the central computing device in MARIN. Among the available devices, the Android-based Samsung Galaxy S III was chosen for computational performance and battery life. As mobile computing technology is evolving at a rapid pace, we anticipate being able to take advantage of the advancing capabilities in this area as they become commercially available, and development within Android provides easy portability across devices.

The computing device serves three main functions: centralizing data collection, providing the user interface, and collecting self-reports and survey data. Sensors included in the final system all utilize Bluetooth or WiFi for data transmission, with the phone serving as host. Real-time data from the sensors is streamed to the computing device, where it is time stamped and recorded in a combined data store. Due to inherent delays in wireless data transmission, it is anticipated that there will be small variations in synchronization of the data from the various sensors. For the supplementary sensors, this should not pose an issue because the time resolution of the measures (heart rate, respiration rate, skin temperature, electrodermal activity) is such that a several millisecond delay in correlating to EEG does not affect the usefulness of the data.

A screen is provided for the subject that shows signal quality for the EEG electrodes, the ECG electrodes, and connectivity to the sensors. This enables the subject to put on the system components and immediately see if any sensors need to be adjusted. The computing device monitors the signal quality and connectivity to the sensors throughout the experiment, and if a problem is detected, the subject is notified by an alert on the computing device and provided with instructions to correct the issue. Due to the simplistic, user-friendly nature of the interface, we anticipate only minimal training will be necessary for users to become proficient with applying and monitoring the system components.

We have also developed a range of applications on the Android platform to enable an observational, multi-aspect measurement approach targeted at building a context to interpret the neural activity related to stress throughout the day. These applications

currently include the main subject-interaction panel, master scheduler, and questionnaires administrator.

The Main Subject-Interaction Panel. A large widget on the home screen of the Android device allows the user to self-report the start and end of pre-defined activities (eating, drinking, meeting, conversing, etc.), as well as unexpected events (startling sounds, equipment adjustment or repositioning, data logging mistakes, etc.). The widget is comprised of ten buttons. These allow the user to: (1) report the start/stop of reading email; (2) report the start/stop of consumption of a caffeinated beverage; (3) report the start/stop of consumption of food; (4) report the start/stop of a conversation; (5) report the start/stop of a meeting; (6) report the start/stop of exercise; (7) report the start/stop of listening to music; (8) view the experiment schedule; (9) access some application settings; and (10) report an incident. A screenshot is depicted in the center of Figure 2.

When certain events (items 1-7) are logged by the subject, short surveys are administered to gain more information about the event. For example, when the caffeine or food buttons are pressed, the user is asked to rate the size of the beverage or meal. Meanwhile, the remaining three buttons serve utilitarian functions for the user to ensure smooth usability, such as providing the ability to view a textual display of the experiment schedule for the day, or allowing the user to change the current subject ID and to activate or deactivate the schedule alarms. Finally, because not all event types can be predicted or classified ahead of time, the report incident button provides the user a way to input a generic text description of any other event.

The Master Scheduler. A background application is responsible for triggering the various alarms and events detailed in the experiment schedule. It sets off an alarm with a textual reminder for each activity. Some examples of reminders would be to start a task, take off the equipment, or fill out some questionnaires. In the case of the questionnaires or certain tasks, the master scheduler starts the relevant app automatically.

The Questionnaire Administrator. An application houses all of the questionnaires for the experiment. Currently, these include a variety of inventories related to stress, e.g. a Visual Analog Scale of Stress (S-VAS) [26], and variables that influence stress, e.g. the NASA Task Loading Index (TLX) [27] and the Pittsburgh Sleep Diary [28].

4 Conclusion

This paper has described a novel, laboratory-grade multi-modal neuroimaging system designed to overcome the obstacles of real-world neuroimaging. A major obstacle this system has addressed is the need for context monitoring that will result in a meaningful interpretation of observed real-world neural signals. This system is particularly suited to the scientific study of stress given how it integrates physiological responses related to the autonomic nervous system with high quality neurological data and subjective measurements. This technology will lead to a better understanding of neural activity in the real world, which ultimately will help develop better neurotechnology.

References

1. Lamond, N., Dawson, D., Roach, G.D.: Fatigue assessment in the field: validation of a hand-held electronic psychomotor vigilance task. Aviat Space Environ. Med. 76, 486–489 (2005)
2. Tedeschi, J.T., Quigley, B.M.: Limitations of laboratory paradigms for studying aggression. Aggression and Violent Behavior 1, 163–177 (1996)
3. Levitt, S.D., List, J.A.: What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? The Journal of Economic Perspectives 21, 153–174 (2007)
4. Selye, H.: The stress of life. McGraw-Hill (1956)
5. Chrousos, G.P.: Stress and disorders of the stress system. Nat. Rev. Endocrinol. 5, 374–381 (2009)
6. McEwen, B.S.: Stress, adaptation, and disease. Allostasis and allostatic load. Ann. N. Y. Acad. Sci. 840, 33–44 (1998)
7. Wilhelm, F.H., Grossman, P.: Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. Biological Psychology 84, 552–569 (2010)
8. Zanstra, Y.J., Johnston, D.W.: Cardiovascular reactivity in real life settings: measurement, mechanisms and meaning. Biol. Psychol. 86, 98–105 (2011)
9. Cohen, S., Janicki-Deverts, D., Miller, G.E.: Psychological stress and disease. JAMA 298, 1685–1687 (2007)
10. Aldwin, C.M., Molitor, N.-T., Spiro, A., Levenson, M.R., Molitor, J., Igarashi, H.: Do Stress Trajectories Predict Mortality in Older Men? Longitudinal Findings from the VA Normative Aging Study. Journal of Aging Research 2011, 1–10 (2011)
11. Lantz, P.M., House, J.S., Mero, R.P., Williams, D.R.: Stress, Life Events, and Socioeconomic Disparities in Health: Results from the Americans' Changing Lives Study. Journal of Health and Social Behavior 46, 274–288 (2005)
12. Staal, M.A.: Stress, cognition, and human performance: A literature review and conceptual framework (2004)
13. Starcke, K., Brand, M.: Decision making under stress: a selective review. Neurosci. Biobehav Rev. 36, 1228–1248 (2012)
14. Dawson, M.E., Schell, A.M., Filion, D.L.: The electrodermal system. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.) Handbook of Psychophysiology, 3rd edn., pp. 159–181. Cambridge University Press, New York (2007)
15. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems 6, 156–166 (2005)
16. Poh, M.Z., Swenson, N.C., Picard, R.W.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Transactions on Biomedical Engineering 57, 1243–1252 (2010)
17. Liao, L.-D., Lin, C.-T.: Novel Trends in Biosensors Used for Electroencephalography Measurements in Neurocognitive Engineering Applications. Journal of Neuroscience and Neuroengineering 1, 32–41 (2012)
18. Hosseini, S.A., Khalilzadeh, M.A.: Emotional stress recognition system using EEG and psychophysiological signals: Using new labelling process of EEG signals in emotional stress state. In: 2010 International Conference on Biomedical Engineering and Computer Science (ICBECS), pp. 1–6 (2010)

19. Hamid, N.H.A., Sulaiman, N., Aris, S.A.M., Murat, Z.H., Taib, M.N.: Evaluation of human stress using EEG Power Spectrum. In: 2010 6th International Colloquium on Signal Processing and Its Applications (CSPA), pp. 1–4 (2010)
20. Sulaiman, N., Taib, M.N., Lias, S., Murat, Z.H., Aris, S.A.M., Mustafa, M., Rashid, N.A.: Development of EEG-based stress index. In: 2012 International Conference on Biomedical Engineering (ICoBE), pp. 461–466 (2012)
21. Sulaiman, N., Taib, M.N., Lias, S., Murat, Z.H., Mustafa, M., Aris, S.A.M., Rashid, N.A.: Electroencephalogram-Based Stress Index. Journal of Medical Imaging and Health Informatics 2, 327–335 (2012)
22. Wilhelm, F.H., Roth, W.T., Sackner, M.A.: The lifeShirt. An advanced system for ambulatory measurement of respiratory and cardiac function. Behav. Modif. 27, 671–691 (2003)
23. Lin, C.-T., Liao, L.-D., Liu, Y.-H., Wang, I.-J., Lin, B.-S., Chang, J.-Y.: Novel Dry Polymer Foam Electrodes for Long-Term EEG Measurement. IEEE Transactions on Biomedical Engineering 58, 1200–1207 (2011)
24. Gramann, K., Gwin, J.T., Ferris, D.P., Oie, K., Jung, T.-P., Lin, C.-T., Liao, L.-D., Makeig, S.: Cognition in action: imaging brain/body dynamics in mobile humans. Rev. Neurosci. 22, 593–608 (2011)
25. Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M.J., Iragui, V., Sejnowski, T.J.: Removing electroencephalographic artifacts by blind source separation. Psychophysiology 37, 163–178 (2000)
26. Cella, D.F., Perry, S.W.: Reliability and concurrent validity of three visual-analogue mood scales. Psychol. Rep. 59, 827–833 (1986)
27. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock, P.A., Meshkati, N. (eds.) Advances in Psychology, pp. 139–183. North-Holland (1988)
28. Monk, T.H., Reynolds, C.F., Kupfer III, D.J., Buysse, D.J., Coble, P.A., Hayes, A.J., Machen, M.A., Petrie, S.R., Ritenour, A.M.: The Pittsburgh Sleep Diary. J. Sleep Res. 3, 111–120 (1994)

Optimal Feature Selection for Artifact Classification in EEG Time Series

Vernon Lawhern[1], W. David Hairston[2], and Kay Robbins[1]

[1] Department of Computer Science, University of Texas-San Antonio, San Antonio, TX 78249 USA
[2] Human Research and Engineering Directorate, Army Research Laboratory, Aberdeen Proving Ground, MD 21005 USA
{Vernon.Lawhern,Kay.Robbins}@utsa.edu,
william.d.hairston4.civ@mail.mil

Abstract. Identifying artifacts or non-brain electrical signals in EEG time series is often a necessary but time-consuming preprocessing step, as many EEG analysis techniques require that the data be artifact free. Because of this, reliable and accurate techniques for automated artifact detection are desirable in practice. Previous research has shown that coefficients obtained from autoregressive (AR) models can be used as feature vectors to classify among several different artifact conditions found in EEG. However, a statistical method for identifying significant AR features has not been presented. In this work we propose a method for determining the optimal AR features that is based on penalized multinomial regression. Our results indicate that the size of the feature vector can be greatly reduced with minimal loss to classification accuracy. The features selected by this algorithm localize to specific channels and suggests a possible BCI implementation with increased computational efficiency than with using all available channels. We also show that the significant AR features produced by this approach correlate to known brain physiological properties.

Keywords: Autoregressive (AR) model, Artifacts, Electroencephalography, classification, feature selection, multinomial regression, penalized regression, machine learning.

1 Introduction

In current EEG, the signal of interest is easily confounded by other biological sources of voltage, often stemming from muscle (EMG) or eye (EOG) movements. Great care is taken in laboratory settings to limit sources of artifacts, such as by having subjects limit any unnecessary movements or actions during the experiment, as these activities may confound the EEG activities of interest. After completing an experiment, researchers still must remove artifacts in EEG signals to obtain a "clean" signal that can be further analyzed. This process often requires manual identification of artifact-contaminated EEG, generally conducted by a panel of experts, which can be tedious and time-consuming, especially for large amounts of data. New applications of EEG

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 326–334, 2013.

are being performed in more complex and realistic environments, where controlling the effects of artifacts is not feasible, such as in the detection of fatigue while driving [1]. Similarly, brain-computer interfaces (BCIs) are being developed for individuals who may have physical disabilities and as a means to improve performance in healthy individuals [2]. In these scenarios, traditional labor-intensive off-line analyses that require extensive computation to remove artifacts are not feasible. Thus, extending applications of EEG to more realistic scenarios will require automated artifact detection methods that are robust to both inter-subject and intra-subject variations.

A common approach for the analysis of EEG signals is autoregressive (AR) modeling. Autoregressive models are linear models that relate signals to their past values. The coefficients of these models can characterize signal properties. Single-channel AR models relate signals to their own values, while multivariate models can model relationships between simultaneously-recorded time series. Useful characteristics of time series can be derived such as ordinary, partial or directed coherence [3, 4] and the direct transfer function (DTF) [5]. AR models are attractive representations in that they are compact and computationally efficient.

One important feature of AR models is that the coefficients are invariant to scaling changes in the data, making AR approaches valuable in EEG analyses. AR modeling has been extensively used in EEG data analysis for feature extraction and classification tasks [6], detection and classification of cardiac arrhythmias [7], and analysis of epilepsy data [8]. AR models have also been used for detecting artifacts in EEG signals. For example, Van de Velde et al [9] used features such as the slope, signal variance and AR model coefficients to classify EEG segments into three artifact categories: None, Moderate and Severe.

Our recent work [10] has shown that AR coefficients can be used alone to classify type-specific artifacts such as eye blinks and jaw movements. This method uses AR coefficients together with a support vector machine (SVM) classifier to distinguish among 8 different artifact conditions. While this is already a relatively efficient method, the high degree of correlation in the signals from neighboring channels and the close relationship of their resulting AR features exhibit a high degree of redundancy if all channels of a high-density cap are included. This suggests substantial room for streamlining the computation and very likely the hardware necessary for data acquisition. Channel elimination requires reliable methods for down-selecting channels, and the high degree of correlation among the features may make traditional feature selection techniques such as AIC (Akaike information criterion) or BIC (Bayesian information criterion) unreliable. There may also be situations where there are many more parameters than samples (the $p >> N$ case), making this an ill-posed problem which cannot be solved using traditional methods. In addition to reasons of analysis, using fewer features has advantages for implementation in natural environments using portable EEG headsets, which usually have many fewer channels than high-density laboratory models. In this environment, processing must be done online and not all channels may be in full contact. Therefore it is valuable to investigate methods that can be used to select only the most important features for EEG signal classification and to understand more clearly how information from different channel loci contribute to classification of different artifact types.

In this paper, we propose a method for determining significant signal features for artifact classification based on regularized multinomial regression. Multinomial regression is an extension of logistic regression, where more than two response classes are present. We use the artifact classes found in [10] as the response levels while using the AR coefficients from EEG channels as the covariates in the model. Since the AR coefficients exhibit a high degree of multi-variable co-linearity, we use an elastic net penalization [11] of the standard maximum likelihood solution to determine the optimal features. This approach has been used successfully in situations where there are many more parameters than samples ($p >> N$) such as in microarray gene expression data and text classification [12]. The high degree of co-linearity can make the matrix inversions needed for standard maximum likelihood unreliable and inaccurate. Our results indicate that a significant reduction in the feature set size is possible without loss in classification accuracy.

2 Experimental Methods

2.1 Experimental Setup

The data used in this study was recorded using a 64-channel Biosemi ActiveTwo System and analyzed in a previous study [10]. A brief summary is given here. A total of seven participants performed a block of artifact-inducing facial and head movements. All provided consent prior to participating, and methods were approved as required by U.S. Army human use regulations [13, 14]. The seven movements included (abbreviations follow): clenching the jaw (JC); moving the jaw vertically (JM); blinking both eyes (EB); moving eyes leftward, then back to center (EL); moving eyes upwards, then back to center (EU); raising and lowering eyebrows (ME); and rotating head side-to-side (as in looking leftward), (RH). All movements were performed sitting in front of a PC screen. The participants were instructed to perform each type of movement 20 times in concert with a consistently occurring tone. A baseline dataset was also recorded for each participant. Participants were told to look straight at the computer screen and to not move excessively in order to minimize muscle artifacts. We extracted 20 epochs of each artifact condition, plus 20 artifact-free epochs from the baseline condition. Our total dataset consisted of 160 epochs, 20 for each of 8 conditions for each of seven participants (see [10] for more details).

3 Statistical Methods

3.1 Autoregressive Models

We use autoregressive (AR) model coefficients as features for artifact classification in EEG. Given a zero mean time series $z_t, t = 1, \ldots, n$, an AR model of order p can be written as:

$$z_t = \sum_{i=1}^{p} A_i z_{t-i} + \epsilon_t \tag{1}$$

where $A_i, i = 1, \ldots, p$ are the AR model coefficients, and $\epsilon_t \sim N(0, \sigma^2)$. The AR model estimates the signal characteristics by modeling the signal compared to the signal in the past p time points. In our analysis EEG channels are modeled individually using a second order AR model, and the AR coefficients are concatenated across channels to form the feature vector used for classification, resulting in a 128-dimensional feature vector. We use the Burg method for fitting the AR coefficients [15].

3.2 Multinomial Regression with Elastic Net Penalization

We treat the classification of artifact signals as a multinomial regression problem, where the artifact classes are the response levels, and the covariates are the AR coefficient features. Let $Y \in \mathbb{R}$ be the response variable (consisting of artifact labels) and $X \in \mathbb{R}^{\mathrm{C}}$ be the vector of AR coefficients (C = 128). Using the notation from [11], the multinomial regression model for the response variable G, having $K > 2$ levels, is:

$$\Pr(G = l|X) = \frac{e^{\beta_{0l} + X^T \beta_l}}{\sum_{k=1}^{K} e^{\beta_{0k} + X^T \beta_k}} \tag{2}$$

where $l = 1, \ldots, K$. We fit this model using regularized multinomial maximum likelihood. Let $p_l(x_i) = \Pr(G = l|x)$ and let $g_i \in \{1,2, \ldots, K\}$ be the i^{th} response. The penalized log-likelihood is:

$$\max_{\{\beta_{0l}, \beta_l\}_1^K \in \mathbb{R}^{K(p+1)}} \left[\frac{1}{N} \sum_{i=1}^{N} \log_{p_{g_i}}(x_i) - \lambda \sum_{l=1}^{K} P_\alpha(\beta_l) \right] \tag{3}$$

where λ is the penalty coefficient and:

$$P_\alpha(\beta_l) = \sum_{j=1}^{C} \left[\frac{1}{2}(1 - \alpha)\beta_{lj}^2 + \alpha|\beta_{lj}| \right] \tag{4}$$

is the *elastic net penalty* [11]. This penalty reduces to the ridge regression penalty when $\alpha = 0$ (the standard l_2 penalty) and the Lasso penalty when $\alpha = 1$ (the standard l_1 penalty). The Lasso penalty is a sparse penalty that forces many of the coefficients to be 0, with a small subset to be nonzero, while the ridge regression penalty shrinks the coefficients of highly correlated variables relative to each other. The parameter α controls the degree of homogeneity among the two penalties. Setting $\alpha = 1 - \epsilon$ for some small ϵ produces a sparse solution similar to Lasso as well as removing irregular behavior caused by a high degree of co-linearity among the covariates. In our analysis we set $\alpha = .99$ as we seek a sparse solution that is robust to high correlations among covariates. We use the GLMNET toolbox for MATLAB [11] to solve for the coefficients. The optimal λ is found by using a grid search and maximizing the percentage of explained deviance (see [11] for more details).

3.3 Bootstrap Model Validation

To verify the significance of the model parameters, we randomly partitioned our data into two sets, a training (60%) and testing (40%) set. The training set is used to fit the regularized multinomial model, while the testing set is used to validate the accuracy of the classification. We used B = 100 bootstrap samples and calculated the average accuracy across all the samples. Note that the significant covariates may change at each bootstrap iteration; therefore, in a separate analysis, the covariates that were significant in at least 75% of the bootstrap iterations were extracted and a ridge regression model ($\alpha = 0$) was used on only these covariates. A ridge regression model was used as high degree of co-linearity may still exist among these covariates.

4 Results

The results of our classification study are shown in Table 1. The first row within each subject grouping denotes the classification accuracies when using all available parameters in the data and using the radial basis function support vector machine (RBF-SVM) that was used in [10] for artifact classification. The results from this classification are taken as the baseline performance, which we compare our current methods against. The average classification accuracy over all subjects is 95.8% +/- 2%. The second row denotes the classification accuracy from the elastic net penalty for the multinomial regression. The average performance in this case is not significantly different than using the full feature vector with SVM (94.7% +/- 2.4%) while using significantly fewer parameters in the model (40.3). This result indicates that the AR feature vector is highly redundant and in fact the majority of features are not necessary to obtain the same classification accuracy. When using only the parameters that appeared in at least 75% of the bootstrap iterations (third row within subject), we see a slight reduction in accuracy of ~4-5%. A Kruskal Wallis ANOVA revealed only minimal evidence of a significant difference in the three classification probabilities ($\chi^2 = 7.48$, $p < .03$). Note that subject 7 saw no decrease in overall performance between the two models, while subjects 3 and 6 saw minimal reduction (3% or less).

Figure 1 shows a channel plot of significant channels for all of the subjects in the analysis. The first plot (top left) denotes the standard configuration of the 64-channel Biosemi System (see Materials and Methods). Channels in red indicate that at least one of the two AR(2) coefficients was significant in at least 75% of bootstrap samples, while channels in blue indicate both the AR(2) coefficients were significant in at least 75% of bootstrap samples. We see that there is some degree of consistency across subjects, with channels located frontally significant, while a few channels around the edge of the cap are also consistently contributing to the discrimination.

Figure 2 shows the classification performance for different criterion percentage values of the bootstrap models. The x-axis value at 0 denotes the classification percentage using the full feature vector (128 parameters) similar to the SVM-only classifier as in [10]. The bootstrap percentage value at 20 indicates that we use the parameters that occur in at least 20% of bootstrap models to build the multinomial regression. The two y-axes denote the percentage of the total number of parameters

used in the model (blue, left side, which varies by percentage criterion) and the resulting overall classification percentage (green, right side). For example, at the bootstrap percentage value of 20% (meaning parameters had to appear in at least 20% of the bootstrap models to be included for analysis), about 40% of the parameters were used (~54 parameters) while achieving a classification percentage of ~93%. While there is a dramatic drop in the percent of parameters remaining in the model, which tapers to a slower decline, we simultaneously see that the accuracy curve (green) remains fairly flat until after the 80% bootstrap percentage value, where a noticeable reduction (to about 83%) occurs.

Table 1. Classification percentages for the elastic net regression models for classifying artifact conditions based on the average of 100 bootstrap models. Values in parentheses denote one standard deviation of the classification percentage. The first row within each subject denotes the average classification probabilities using all available parameters and using the SVM for classification. The second row denotes the average classification probability using elastic net penalization method, while the third row denotes the average classification probability only using parameters that were significant in >75% of the bootstrap models and using ridge regression to fit the multinomial model. The *P* column in the first row of each subject denotes the average number of significant parameters. Mean = average accuracy for all movements, JC = Jaw Clench, JM = Jaw Movement, EB = Eye Blink, EL = Eye Left Movement, EU = Eye Up Movement, ME = Move Eyebrows, RH = Rotate Head.

Subj	*P*	*Mean*	*JC*	*JM*	*EB*	*EL*	*EU*	*ME*	*RH*	*None*
1	128	96(1.8)	99(2.8)	99(2.8)	87(10.3)	93(6.3)	94(7.5)	100(0)	98(4.5)	96(7.1)
	38.7	95(2.4)	99 (1.2)	95(6.1)	89(10.9)	90(10.0)	98(5.1)	100(0)	97(5.7)	92(10.7)
	26	91(3.6)	94(10.2)	88(10.7)	91(7.9)	86(10.5)	97(5.8)	92(10.3)	93(9.8)	85(13.8)
2	128	93(2.4)	99(2.8)	92(7.3)	100(0)	97(5.5)	86(10.6)	89(7.3)	88(10.3)	89(12.3)
	49.7	90(3.3)	99 (3.2)	84(13.8)	98(3.7)	92(8.3)	90(10.9)	89(8.5)	75(13.3)	84(12.3)
	24	86(3.6)	99(2.1)	77(14.4)	98(4.3)	90(11.8)	81(14.5)	86(10.3)	84(13.3)	75(14.7)
3	128	97(2.3)	100(0)	89(10.1)	100(0)	100(0)	92(8.3)	100(0)	97(6.8)	96(6.1)
	35	98(1.6)	95(7.3)	94(9.4)	99(2.1)	100(0)	100(0)	99(1.2)	96(5.8)	100(0)
	19	95(2.1)	99(2.1)	90(9.4)	99(2.4)	98(5.2)	84(8.6)	98(4.2)	95(6.4)	94(6.5)
4	128	94(3.6)	100(0)	99(2.8)	98(4.5)	88(14.5)	85(16.5)	99(2.8)	94(11.8)	81(11.1)
	37.9	94(2.8)	99 (1.2)	100(0)	96(5.7)	90(10.3)	91(9.8)	98(4.7)	87(10.2)	91(10.1)
	20	89(3.6)	97(7.8)	99(2.4)	96(5.5)	83(12.1)	87(11.3)	91(9.7)	78(13.6)	79(14.1)
5	128	97(2.1)	100(0)	100(0)	99(2.8)	84(10.8)	93(10.2)	100(0)	100(0)	100(0)
	41.0	95(3.4)	100(0)	99(3.8)	100(0)	86(11.5)	86(13.5)	96(5.9)	95(10.1)	99(2.7)
	23	90(3.6)	96(6.1)	95(7.8)	99(1.7)	67(17.1)	92(9.0)	83(11.5)	93(8.2)	90(12.1)
6	128	97(1.9)	95(6.2)	99(2.8)	98(5.1)	96(7.1)	94(6.3)	97(5.5)	99(3.8)	100(0)
	41.4	96(2.5)	98(5.3)	90(11.4)	96(5.9)	99(2.9)	96(7.7)	99(3.6)	93(9.4)	100(0)
	26	93(2.7)	91(7.4)	94(6.9)	95(6.2)	97(5.4)	96(6.0)	84(11.2)	93(10.8)	96(5.9)
7	128	98(1.7)	98(5.1)	95(7.4)	93(6.2)	99(2.8)	100(0)	97(5.5)	100(0)	100(0)
	38.7	95(2.4)	99 (1.2)	95(6.1)	89(10.9)	90(10.0)	98(5.1)	100(0)	97(5.7)	92(10.7)
	27	95(2.4)	97(5.2)	94(7.2)	84(12.3)	95(6.2)	91(9.2)	97(6.1)	100(0)	100(0)

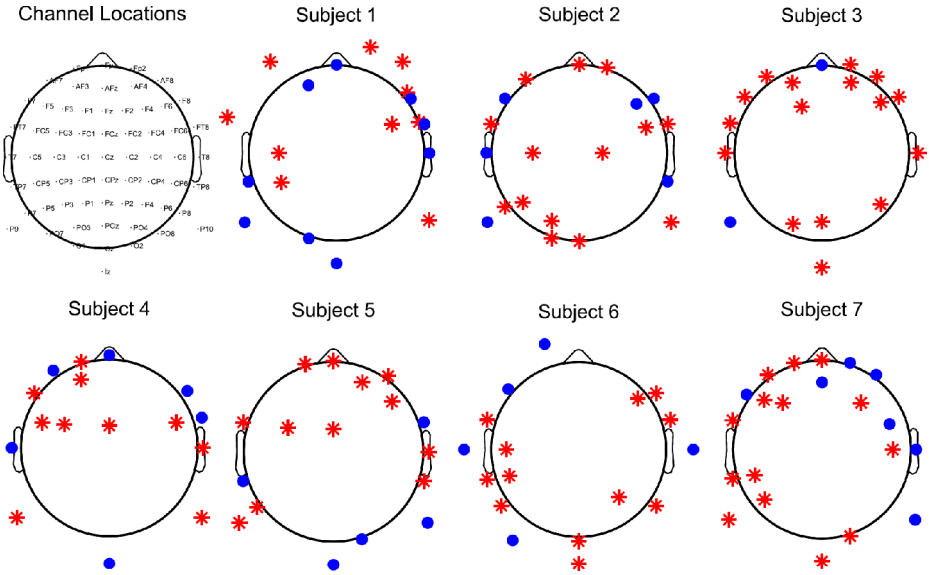


Fig. 1. Plot of significant channels for all subjects in the study. The first plot depicts the 10-20 channel orientation of a 64-channel Biosemi System. Channels with red stars indicate that at least one of the two AR(2) coefficients was significant in at least 75% of bootstrap samples, while channels with blue circles indicate both AR(2) coefficients were significant at this same criterion.

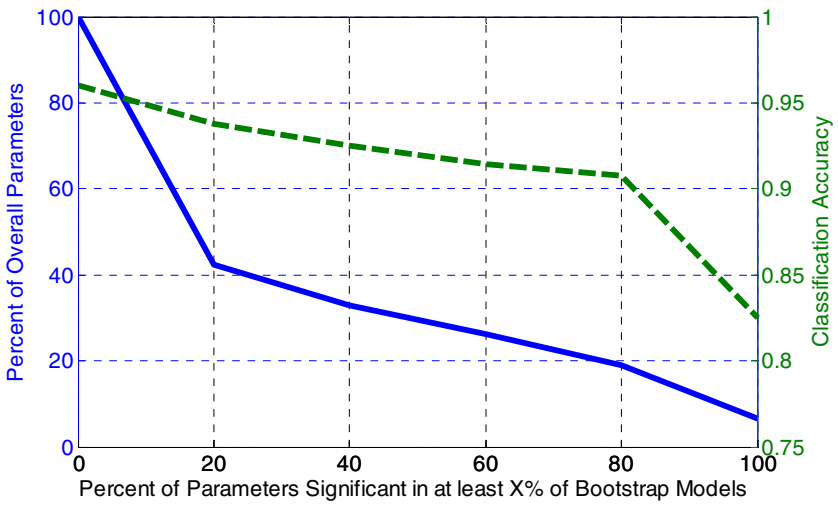


Fig. 2. Plot of the average percentages of overall parameters and the classification percentage for different percentage of parameters observed in bootstrap models. The dashed green line denotes the classification accuracy, while the solid blue line denotes the percent of overall parameters used in the model.

5 Conclusion

In this paper we have proposed a method for down-selecting the appropriate features necessary for accurate discrimination of EEG artifacts based on elastic net penalized regression models. The elastic net penalty applied to multinomial regression can effectively handle the high correlations and redundancy in the AR parameters and appears to be an effective general approach for feature selection in EEG analysis. In our analysis, using the elastic net penalty with multinomial regression effectively reduced the number of parameters by 60% without any loss in classification accuracy. The overall classification accuracy remained above 90% until we restricted the number of parameters to less than 20% of the overall parameters available (Fig 2). This indicates that a significant computational savings could be achievable if implemented in a BCI system. For example, data streamlining is critical in new wireless EEG headsets, where transmission bandwidth is limited by power. Although the high variability observed across subjects might limit the possibility of physically tailoring the channel locations to a specific user, one possible scheme might be to only record and broadcast data from the channels previously established to be most meaningful for that individual. Potential applications of this approach include monitoring subjects for artifact instances such as eye blink frequency and duration for detecting lapses in attention during experiments [16].

The results derived from artifact classification by the regularized multinomial regression are corroborated by known brain physiological properties. For example, there were many frontal channels identified as being highly significant, which is expected given that these channels exhibit eye movement artifacts the most strongly. Meanwhile, there were also many significant channels located around the edges of the cap, while the majority of those in the center are less likely to significantly contribute to the discrimination. One possible reason for this is that muscle activations from the rotate head (RH) condition are picked up by the channels located near the neck. Channels near the ears are also significant in many subjects, as these channels are located near the jawline and pick up jaw clench and jaw movement artifacts. A few channels located at the top are most likely contributing to the model of the baseline condition, as these channels are minimally impacted by artifacts.

Acknowledgments. We thank Scott Kerick and Kaleb McDowell of the Army Research Laboratory for helpful discussions and for help with data collection. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Lin, C.-T., Chang, C.-J., Lin, B.-S., Hung, S.-H., Chao, C.-F., Wang, I.-J.: A Real-Time Wireless Brain-Computer Interface System for Drowsiness Detection. IEEE Transactions on Biomedical Circuits and Systems 4, 214–222 (2010)
2. Lance, B.J., Kerick, S.E., Ries, A.J., Oie, K.S., McDowell, K.: Brain-Computer Interface Technologies in the Coming Decades. Proceedings of the IEEE 100, 1585–1599 (2012)
3. Baccalá, L.A., Sameshima, K.: Partial directed coherence: a new concept in neural structure determination. Biol. Cybern. 84, 463–474 (2001)
4. Möller, E., Schack, B., Arnold, M., Witte, H.: Instantaneous multivariate EEG coherence analysis by means of adaptive high-dimensional autoregressive models. J. Neurosci. Methods 105, 143–158 (2001)
5. Franaszczuk, P.J., Bergey, G.K., Kamiński, M.J.: Analysis of mesial temporal seizure onset and propagation using the directed transfer function method. Electroencephalography and Clinical Neurophysiology 91, 413–427 (1994)
6. Anderson, C.W., Stolz, E.A., Shamsunder, S.: Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Transactions on Biomedical Engineering 45, 277–286 (1998)
7. Ge, D., Srinivasan, N., Krishnan, S.M.: Cardiac arrhythmia classification using autoregressive modeling. BioMedical Engineering OnLine 1, 5 (2002)
8. Übeyli, E.D.: Least squares support vector machine employing model-based methods coefficients for analysis of EEG signals. Expert Systems with Applications 37, 233–239 (2010)
9. Van de Velde, M., Ghosh, I.R., Cluitmans, P.J.M.: Context related artefact detection in prolonged EEG recordings. Computer Methods and Programs in Biomedicine 60, 183–196 (1999)
10. Lawhern, V., Hairston, W.D., McDowell, K., Westerfield, M., Robbins, K.: Detection and classification of subject-generated artifacts in EEG signals using autoregressive models. J. Neurosci. Methods 208, 181–189 (2012)
11. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1 (2010)
12. Zhu, J., Hastie, T.: Classification of gene microarrays by penalized logistic regression. Biostatistics 5, 427–443 (2004)
13. U.S. Department of Defense, Office of the Secretary of Defense: Code of federal regulations, protection of human subjects. Government Printing Office. 32 CFR 19 (1999)
14. U.S. Department of the Army: Use of volunteers as subjects of research. Government Printing Office. AR 70-25 (1990)
15. Schlögl, A.: A comparison of multivariate autoregressive estimators. Signal Processing 86, 2426–2429 (2006)
16. Kim, Y.S., Baek, H.J., Kim, J.S., Lee, H.B., Choi, J.M., Park, K.S.: Helmet-based physiological signal monitoring system. Eur. J. Appl. Physiol. 105, 365–372 (2009)

Towards a Hybrid P300-Based BCI Using Simultaneous fNIR and EEG

Yichuan Liu[1,2], Hasan Ayaz[1,2], Adrian Curtin[1,2], Banu Onaral[1,2], and Patricia A. Shewokis[1,2,3]

[1] School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, PA 19104, USA
[2] Cognitive Neuroengineering and Quantitative Experimental Research (CONQUER) Collaborative, Drexel University, Philadelphia, PA 19104, USA
[3] Nutrition Sciences Department, College of Nursing and Health Professions, Drexel University, Philadelphia, PA 19102, USA
Yichuan.Liu565@drexel.edu

Abstract. Next generation brain computer interfaces (BCI) are expected to provide robust and continuous control mechanism. In this study, we assessed integration of optical brain imaging (fNIR: functional near infrared spectroscopy) to a P300-BCI for improving BCI usability by monitoring cognitive workload and performance. fNIR is a safe and wearable neuroimaging modality that tracks cortical hemodynamics in response to sensory, motor, or cognitive activation. Eight volunteers participated in the study where simultaneous EEG and 16 optode fNIR from anterior prefrontal cortex were recorded while participants engaged with the P300-BCI for spatial navigation. The results showed a significant response in fNIR signals during high, medium and low performance indicating a positive correlation between prefrontal oxygenation changes and BCI performance. This preliminary study provided evidence that the performance of P300-BCI can be monitored by fNIR which in turn can help improve the robustness of the BCI classification.

Keywords: BCI, P300, fNIR, Performance, Optical brain imaging, EEG.

1 Introduction

A brain-computer interface (BCI) decodes neurophysiological signals from the brain for direct controlling an external device without the brain's normal communication pathway of peripheral nerves and muscles. Electroencephalography (EEG) is by far the most studied technology for non-invasive BCI signal acquisition [1-3]. Apart from EEG, variant types of signal acquisition methods such as Magnetoencephalography (MEG) [4], functional near-infrared spectroscopy (fNIR) [5-9] and functional magnetic resonance imaging (fMRI) [10, 11] has been proposed to be applied in BCI. More recently, several studies showed that utilizing multimodal neuroimaging has the potential to enhance BCI performance [12-16]. These BCIs were generally referred to as hybrid BCIs in the literature.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 335–344, 2013.

In this pilot study, our aim was to investigate combining fNIR and EEG for enhancing a P300 based BCI. P300 is an event-related potential usually elicited by the oddball paradigm. A typical P300-BCI show to the user sequences of stimulus and the user's task is to identify the infrequent occurrence of the target stimulus. Since early works of Farwell and Donchin in the 1980s [2], substantional progress has been made for enhancing the capability of the P300-BCI [See Mak and McFarland [17] for a detail review]. Despite the volume and depth of work conducted in this area, to our best knowledge, to date no study has been done to investigate the possible benefit of combining fNIR and EEG in a P300-BCI.

fNIR is an optical brain imaging technology for monitoring the changes in the concentration of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) in the cortex. Typically, neuronal activities in the active area of the cortex would eventually cause an overabundance of local blood oxygenation result from a mechanism known as neurovascular coupling [18]. Coyle et al in 2004 proposed using a single channel NIR for developing a mind-switch [5]. Sitaram et al [6] demonstrated multichannel optical BCI for hemodynamic pattern classification for motor imagery. Ayaz et al in 2007 proposed using fNIR cognitive tasks for on/off switch [7] while Fazli et al in 2012 showed that combining EEG and NIR can significantly improve motor imagery BCI [19]. The same group also showed that fNIR can serve as a predictor for the performance of EEG-based motor imagery BCI [20].

Recently, several studies investigated predicting the between-subject performance (or aptitude) of P300-BCI [21, 22] based on EEG predictors. In [21], the within-subject effects were also investigated but no significant predictors were found. Predicting the within-subject performance is of particular interest because it may provide information for generating more robust BCI classifiers. In [20], fNIR predicted motor imagery BCI performance was used for generating a meta-classifier which enhanced classification accuracy. In this study, we propose using a prefrontal cortex based fNIR for monitoring within-subject performance of a P300-BCI. It has been established in fMRI studies that the BOLD signal is associated with varies event-related tasks [23-25]. Previous work also suggested that the prefrontal cortex is associated with the level of alertness and attention [26-28] which can affect BCI performance. A fNIR study by our group showed that prefrontal activations were correlated with the performance of an n-back task [29]. The aforementioned evidence suggests a possible correlation between prefrontal activation and P300-BCI performance. For testing the hypothesis, prefrontal fNIR was recorded while subjects were using a spatial navigation P300-BCI that we proposed previously [30, 31]. Our preliminary results show that the subject-wise performance of P300-BCI may be monitored by prefrontal fNIR recording.

2 Materials and Methods

2.1 Participants

Eight right-handed healthy students from local universities participated in this study. The participants included 5 males, 3 females and ages between 22 to 26 years. All

participants did not have prior experience with BCI and gave written informed consent approved by the institutional review board of Drexel University for the experiment. The first three of the participants were excluded from the analysis due to technical issues such as missing synchronization markers or poor signal quality.

2.2 Experiment Setup

The experimental setup was based on a spatial navigation P300-BCI we proposed in [31]. Subjects sat conformably inside a faraday cage. There were two monitors: one was the stimulus presentation monitor placed approximately 30 inches in front of subject on a desk for displaying the P300 BCI matrix; the other was the environment monitor placed on the left hand side of the stimulus presentation monitor for displaying the 3D virtual maze using MazeSuite software [32, 33] (Drexel University). EEG was recorded using Neuroscan Nuamp amplifier and 32 channel EEG cap at 250Hz sampling rate from 9 locations according to 10-20 international system: FCz, Cz, CP3, CPz, CP4, P3, Pz, P4, and Oz. The BCI2000 platform [34] was used for stimuli display and EEG data recording. For online and offline signal processing, MATLAB was used. Prefrontal cortex hemodynamic response was recorded using a16-channel continuous wave fNIR system developed at Drexel University [35] and manufactured by fNIR Device LLC. The sampling rate for fNIR was 2Hz. The COBI Studio Software [32] (Drexel University), which was installed on another desktop, was used for fNIR data recording. Triggers were sent from BCI2000 P3Speller application module to COBI Studio using a serial port periodically for synchronizing the two data streams for offline analysis. Fig.1 shows an overview of the experimental setup.

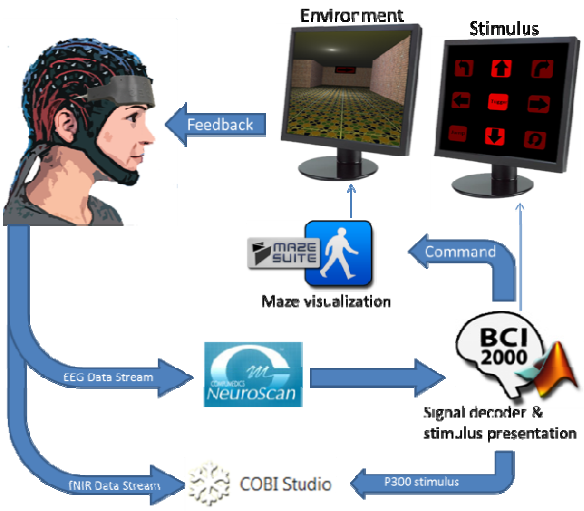


Fig. 1. Experiment setup overview

2.3 Protocol

The visual presentation of P300-BCI included a 3 by 3 matrix of spatial navigation icons (turn left, move forward, turn right, strafe left, trigger, strafe right, jump, move backward and look around, see Fig. 2 Left). During stimulus presentation, row and columns of the matrix were intensified in pseudo random order. The stimulus duration was 80ms and the inter-stimulus interval (ISI) was 160ms. A *sequence* of stimulus included six stimuli – each row and column was intensified exactly once. A *run* included ten sequences at the end of which a command would be outputted from the BCI to the maze.

The experiment included two parts: Part 1 and Part 2. Part 1 was for collecting EEG data in order to calibrate the P300-BCI. It included 24 runs. Before the start of a run, visual instruction was given to the subjects indicating which icon they should attempt to choose by counting the number of times it flashed. For each run, after the end of stimulus presentation, a keyboard with 10 buttons was shown for the subjects to manually record the icon they attempted to choose through activation of the BCI(see Fig. 2 right).

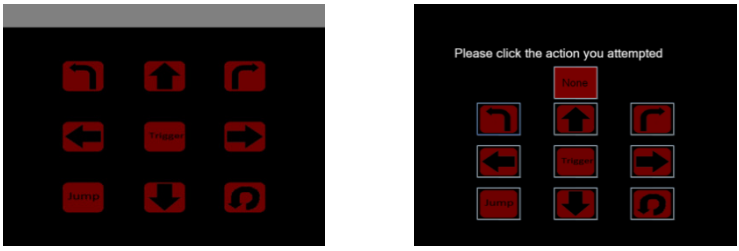


Fig. 2. Left: The 3×3 P300 BCI matrix used in this study. Right: Keyboard shown to the subject.

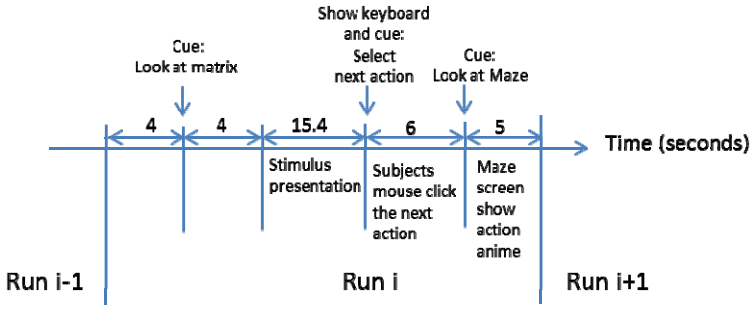


Fig. 3. Time line for a run

Part 2 was for the subjects to navigate freely in 3D virtual mazes using the BCI. Each subject navigated a small mirror maze eight times with possibly different starting points. Fig. 3 shows the timeline of a run in Part 2. The subjects started by looking at the maze screen to decide their next action. A cue 'Look at matrix' would then show on the maze screen asked the subjects to turn their attention to the stimulus

presentation screen. After that, the 10 sequences of stimulus for generating P300 response would be shown at the end of which a keyboard (see Fig. 2. right) was displayed on the same screen for the subjects to record their intended actions in this run. Finally, a cue would show to let the subjects turn their attention back to the maze screen in order to see the maze action animation such as moving forward corresponding to the command output by the BCI.

2.4 Data Processing and Analysis

P300 BCI Classification. Raw EEGs were band pass filtered from 0.5 to 12 Hz and downsampled to 36 Hz. A stepwise linear discriminant analysis (SWLDA) was applied to distinguish target from non-target stimulus based on the EEG amplitudes from 0 to 800ms after the onset of a stimuli. The data collected in Part 1 was used to determine the weights for the classifiers which were then applied to predict the data collected in Part 2.

Performance Criterion for P300 BCI. The performance criterion adopted was the single sequence prediction accuracy (*SeqAcc*) for each run. Target icons were first predicted using the EEG data of each single sequence (note that for a single sequence, each row and column intensified only once). The prediction accuracies for each run were then calculated. Since each run included 10 sequences, this is an ordinal variable with 11 levels of measurement (i.e. from 0 to 1 with 0.1 increments). This criterion gives a finer resolution of the performance and reduced the ceiling effect compared to a simple dichotomous variable indicating whether or not the target of the run has been correctly predicted. Table 1 shows the average and standard deviation of *SeqAcc* for each subject. It can be seen that for subject 4 and 7, their target icon prediction accuracy was the same 100% but *SeqAcc* revealed that the signal quality for subject 7 (*SeqAcc*=0.88±0.12) was much better then subject 4 (*SeqAcc*=0.48±0.19).

fNIR Processing. fNIR signals were first low-pass filtered at 0.1Hz. An automatic artifact detection algorithm, sliding window motion artifact rejection (smar) was employed for eliminating saturation and motion artifact containing segments [35, 36]. Oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) changes were calculated for each P300-BCI run from 0-15s using a local rest period as baseline. To further reduce noise, spatial averaging was performed for both left and right hemisphere separately by averaging the channels located at the left and right hemisphere respectively.

3 Results

P300-BCI Classification. Table 1 listed the sample sizes, target icon prediction accuracy and *SeqAcc* in Part 2 for each subject. Raw sample sizes are different for each subject due to possibly different paths taken during maze navigation, the additional number of runs required for correcting the BCI mistakes and the ratio of rejected run.

Table 1. Sample sizes, target icon prediction accuracy and the Avg. and Std. of *SeqAcc* for each participant

Subject	4	5	6	7	8
Run #	34	81	49	16	58
Accuracy	1.00	0.89	0.78	1.00	0.86
SeqAcc Mean±SD%	0.48±0.19	0.44±0.18	0.38±0.16	0.88±0.12	0.42±0.20

fNIR Results. Fig. 4 shows the grand average fNIR responses for low performance runs and high performance runs during P300 matrix stimulus presentation periods. Each P300 BCI run was categorized into either the low performance group or high performance group subject-wise according to the following criterion:

$$G_{ij} = \begin{cases} High, & if\ S_{ij} > \tilde{S}_i \\ Low, & if\ S_{ij} \le \tilde{S}_i \end{cases} \quad \begin{matrix} i = 4,5,6,7,8 \\ j = 1,2,\dots,n_j \end{matrix}$$

Where S_{ij} is the *SeqAcc* for run j of subject i. $\tilde{S}_i$ is the median *SeqAcc* for subject i. n_j is the number of run for subject i.

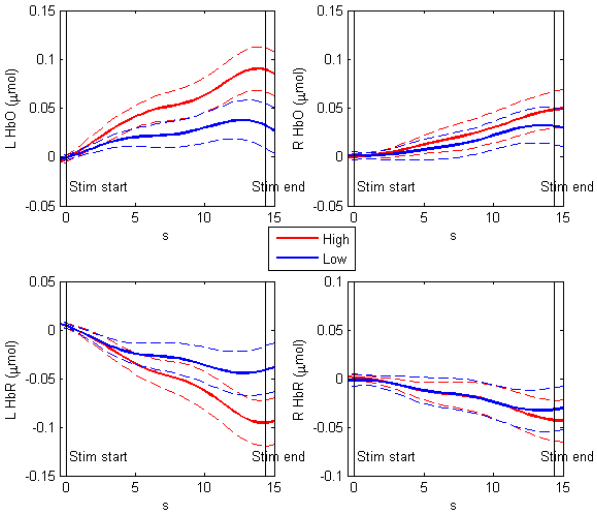


Fig. 4. Grand average fNIR for high performance and low performance runs. The left and right figures show HbO and HbR for left and right hemisphere, respectively. Dash lines stand for standard error of the mean (SEM).

It can be seen that for both left and right hemispheres, HbO was increasing and HbR was decreasing relative to the baseline, consistent with higher activation for prefrontal cortex during the BCI task period. However, for high performance runs, HbO increased (and HbR decreased) at a greater rate compared to low performance runs. Additionally, this phenomenon was more significant for left hemisphere.

Next, the effect of P300-BCI performance on the left hemisphere fNIR response was analyzed. The performances of each BCI run were first categorized into high, medium and low subject-wise according to following criteria:

$$G_{ij} = \begin{cases} High, & if\ S_{ij} > \tilde{S}_i + 0.1 \\ Medium, & otherwise \\ Low, & if\ S_{ij} < \tilde{S}_i - 0.1 \end{cases} \qquad \begin{matrix} j = 1,2,\dots,n_j \\ i = 4,5,6,7,8 \end{matrix}$$

Where $S_{ij}$is the *SeqAcc* for run j of subject i. n_j is the number of run for subject i. $\tilde{S}_i$ is the median *SeqAcc* for subject i.

The fNIR responses were normalized subject-wise before analysis. To reduce sample size, fNIR responses for every three seconds from 0 to 15s were averaged which gives an ordinal time variable with 5 levels. Linear mixed models revealed significant fixed effects of performance ($F_{(2,665.2)}$=7.856, p<0.001; $F_{(2,809.4)}$=10.484, p<0.001) and time ($F_{(4,737.0)}$=10.55, p<0.001; $F_{(4,736.0)}$=11.621, p<0.001) for left hemisphere HbO and HbR, respectively. Bonferroni *post hoc* pairwise comparisons revealed a significantly higher left HbR for low performance runs compared to medium and high performance runs. Conversely, low performance runs have lower left hemisphere HbO levels relative to the medium and high performance runs. Fig. 5 shows the grand average left hemisphere HbO and HbR for the three performance groups.

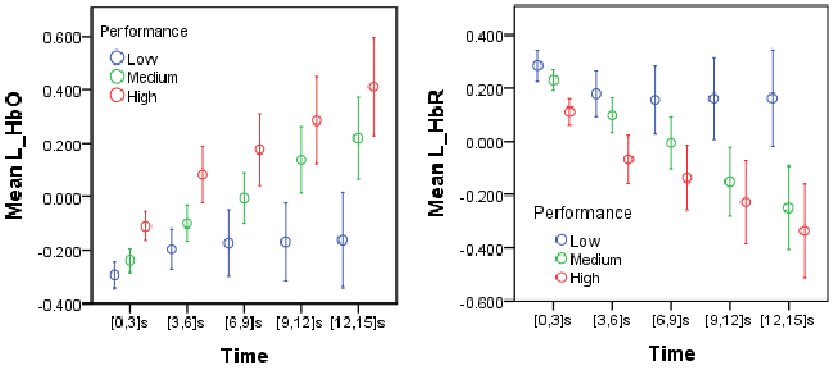


Fig. 5. Grand average left hemisphere HbO (left) and HbR (right) for low, medium and high performance P300-BCI runs. Error bar: standard error (SE).

4 Discussion

In this preliminary study, significant differences in prefrontal activations were found across three levels of within-subject P300-BCI performance. We showed that lower P300-BCI performance was associated with lower level of prefrontal activations, indicating a possible positive correlation between prefrontal activation and BCI performance. Interestingly, Halder et al. in 2011 [37] showed that the performance of a sensorimotor-rhythm BCI is positively correlated with the prefrontal activation. Generally, operating a BCI requires subjects to concentrate on the mental task. Prefrontal

areas, specifically the dorsolateral prefrontal cortex, are associated with attention [38, 39]. Hence, the differences in prefrontal activation across performance levels may be partly due to the different concentration levels during the task periods consistent with our previous results [35, 40].

Despite the encouraging results, more subjects and larger sample sizes are needed for validation. In addition, future studies would benefit from identification of low performance P300-BCI runs to inform a classifier which can help improve the robustness and usability of the BCI. An interesting question is whether some key P300-BCI features such as the amplitude and latency of the P3 and N2 components are correlated with the prefrontal activations. Being able to partially observe the change of these components across time may help adapting the covariate shift due to factors such as alertness and fatigue.

Acknowledgement. This study is made possible in part by a research award from the Intel Corporation and National Science Foundation (NSF) grant IIS:1065471. The content of the information herein does not necessarily reflect the position or the policy of the sponsors and no official endorsement should be inferred.

References

1. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Clinical Neurophysiology 113, 767–791 (2002)
2. Farwell, L.A., Donchin, E.: Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and clinical Neurophysiology 70, 510–523 (1988)
3. Pfurtscheller, G., Neuper, C.: Motor imagery and direct brain-computer communication. Proceedings of the IEEE 89, 1123–1134 (2001)
4. Mellinger, J., Schalk, G., Braun, C., Preissl, H., Rosenstiel, W., Birbaumer, N., Kübler, A.: An MEG-based brain-computer interface (BCI). Neuroimage 36, 581 (2007)
5. Coyle, S., Ward, T., Markham, C., McDarby, G.: On the suitability of near-infrared (NIR) systems for next-generation brain–computer interfaces. Physiological Measurement 25, 815 (2004)
6. Sitaram, R., Zhang, H., Guan, C., Thulasidas, M., Hoshi, Y., Ishikawa, A., Shimizu, K., Birbaumer, N.: Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface. NeuroImage 34, 1416–1427 (2007)
7. Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., Onaral, B.: Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy, pp. 342–345. IEEE (2007)
8. Limongi, T., Di Sante, G., Ferrari, M., Quaresima, V.: Detecting mental calculation related frontal cortex oxygenation changes for brain computer interface using multi-channel functional near infrared topography. International Journal of Bioelectromagnetism 11, 86–90 (2009)

9. Ayaz, H., Shewokis, P., Bunce, S., Schultheis, M., Onaral, B.: Assessment of cognitive neural correlates for a functional near infrared-based brain computer interface system. Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience, 699–708 (2009)
10. Weiskopf, N., Mathiak, K., Bock, S.W., Scharnowski, F., Veit, R., Grodd, W., Goebel, R., Birbaumer, N.: Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI). IEEE Transactions on Biomedical Engineering 51, 966–970 (2004)
11. Yoo, S.S., Fairneny, T., Chen, N.K., Choo, S.E., Panych, L.P., Park, H.W., Lee, S.Y., Jolesz, F.A.: Brain-computer interface using fMRI: spatial navigation by thoughts. Neuroreport 15, 1591–1595 (2004)
12. Ferrez, P.W., Millán, J.R.: Simultaneous real-time detection of motor imagery and error-related potentials for improved BCI accuracy. In: Proceedings of the 4th International Brain-Computer Interface Workshop, pp. 197–202 (2008)
13. Allison, B., Brunner, C., Kaiser, V., Müller-Putz, G., Neuper, C., Pfurtscheller, G.: Toward a hybrid brain–computer interface based on imagined movement and visual attention. Journal of Neural Engineering 7, 026007 (2010)
14. Pfurtscheller, G., Allison, B.Z., Bauernfeind, G.N., Brunner, C., Solis Escalante, T., Scherer, R., Zander, T.O., Mueller-Putz, G., Neuper, C., Birbaumer, N.: The hybrid BCI. Frontiers in Neuroscience 4 (2010)
15. Rebsamen, B., Burdet, E., Zeng, Q., Zhang, H., Ang, M., Teo, C.L., Guan, C., Laugier, C.: Hybrid P300 and mu-beta brain computer interface to operate a brain controlled wheelchair. In: Proceedings of the 2nd International Convention on Rehabilitation Engineering & Assistive Technology, pp. 51-55. Singapore Therapeutic, Assistive & Rehabilitative Technologies (START) Centre, Bangkok, Thailand (2008)
16. Müller-Putz, G.: Hybrid brain-computer interfaces: current state and future directions. PPT 55, 923–929 (2011)
17. Mak, J., Arbel, Y., Minett, J., McCane, L., Yuksel, B., Ryan, D., Thompson, D., Bianchi, L., Erdogmus, D.: Optimizing the P300-based brain–computer interface: current status, limitations and future directions. Journal of Neural Engineering 8, 025003 (2011)
18. Bunce, S.C., Izzetoglu, M., Izzetoglu, K., Onaral, B., Pourrezaei, K.: Functional near-infrared spectroscopy. IEEE Engineering in Medicine and Biology Magazine 25, 54–62 (2006)
19. Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., Blankertz, B.: Enhanced performance by a hybrid NIRS–EEG brain computer interface. Neuroimage 59, 519–529 (2012)
20. Fazli, S., Mehnert, J., Steinbrink, J., Blankertz, B.: Using NIRS as a predictor for EEG-based BCI performance. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4911–4914 (2012)
21. Mak, J.N., McFarland, D.J., Vaughan, T.M., McCane, L.M., Tsui, P.Z., Zeitlin, D.J., Sellers, E.W., Wolpaw, J.R.: EEG correlates of P300-based brain–computer interface (BCI) performance in people with amyotrophic lateral sclerosis. Journal of Neural Engineering 9, 026014 (2012)
22. Halder, S., Hammer, E.M., Kleih, S.C., Bogdan, M., Rosenstiel, W., Birbaumer, N., Kübler, A.: Prediction of Auditory and Visual P300 Brain-Computer Interface Aptitude. PLOS One 8, e53513 (2013)
23. Bledowski, C., Prvulovic, D., Goebel, R., Zanella, F.E., Linden, D.E.J.: Attentional systems in target and distractor processing: a combined ERP and fMRI study. Neuroimage 22, 530–540 (2004)

24. Mccarthy, G., Luby, M., Gore, J., Goldman-Rakic, P.: Infrequent events transiently activate human prefrontal and parietal cortex as measured by functional MRI. Journal of Neurophysiology 77, 1630–1634 (1997)
25. Horovitz, S.G., Skudlarski, P., Gore, J.C.: Correlations and dissociations between BOLD signal and P300 amplitude in an auditory oddball task: a parametric approach to combining fMRI and ERP. Magnetic Resonance Imaging 20, 319–325 (2002)
26. Thomas, M., Sing, H., Belenky, G., Holcomb, H., Mayberg, H., Dannals, R., Wagner, J., Thorne, D., Popp, K., Rowland, L.: Neural basis of alertness and cognitive performance impairments during sleepiness. I. Effects of 24 h of sleep deprivation on waking human regional brain activity. Journal of Sleep Research 9, 335–352 (2008)
27. Moller, H.J., Rizzo, A.A., Mikulis, D.J.: Prefrontal cortex activation mediates cognitive reserve alertness and attention in the Virtual Classroom: preliminary fMRI findings and clinical implications. In: Virtual Rehabilitation, pp. 146–150 (2007)
28. Fuster, J.M.: The prefrontal cortex. Academic Press (2008)
29. Ayaz, H., Bunce, S., Shewokis, P., Izzetoglu, K., Willems, B., Onaral, B.: Using Brain Activity to Predict Task Performance and Operator Efficiency. In: Zhang, H., Hussain, A., Liu, D., Wang, Z. (eds.) BICS 2012. LNCS, vol. 7366, pp. 147–155. Springer, Heidelberg (2012)
30. Liu, Y., Ayaz, H., Curtin, A., Shewokis, P.A., Onaral, B.: Detection of attention shift for asynchronous P300-based BCI. In: Proc. IEEE Eng. Med. Biol. Soc., pp. 4724–4727 (2012)
31. Curtin, A., Ayaz, H., Liu, Y., Shewokis, P.A., Onaral, B.: A P300-based EEG-BCI for Spatial Navigation Control. In: Conf. Proc. IEEE Eng.Med. Biol. Soc., pp. 3841–3844 (2012)
32. Ayaz, H., Shewokis, P.A., Curtin, A., Izzetoglu, M., Izzetoglu, K., Onaral, B.: Using MazeSuite and functional near infrared spectroscopy to study learning in spatial navigation. J. Vis. Exp., e3443 (2011)
33. Ayaz, H., Allen, S.L., Platek, S.M., Onaral, B.: Maze Suite 1.0: a complete set of tools to prepare, present, and analyze navigational and spatial cognitive neuroscience experiments. Behavior Research Methods 40, 353–359 (2008)
34. Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R.: BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Transactions on Biomedical Engineering 51, 1034–1043 (2004)
35. Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B.: Optical brain monitoring for operator training and mental workload assessment. Neuroimage 59, 36–47 (2012)
36. Ayaz, H., Izzetoglu, M., Shewokis, P.A., Onaral, B.: Sliding-window motion artifact rejection for functional near-infrared spectroscopy. In: Annual International Conf. on Engineering in Medicine and Biology Society (EMBC), pp. 6567–6570. IEEE (2010)
37. Halder, S., Agorastos, D., Veit, R., Hammer, E., Lee, S., Varkuti, B., Bogdan, M., Rosenstiel, W., Birbaumer, N., Kübler, A.: Neural mechanisms of brain–computer interface control. NeuroImage 55, 1779–1790 (2011)
38. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. Annual Review of Neuroscience 24, 167–202 (2001)
39. Posner, M.I., Rothbart, M.K.: Research on attention networks as a model for the integration of psychological science. Annu. Rev. Psychol. 58, 1–23 (2007)
40. Ayaz, H., Shewokis, P.A., İzzetoğlu, M., Çakır, M.P., Onaral, B.: Tangram solved? Prefrontal cortex activation analysis during geometric problem solving. In: 34th Annual International IEEE EMBS Conference, pp. 4724–4727. IEEE (2012)

A Novel Method for Single-Trial Classification in the Face of Temporal Variability

Amar R. Marathe, Anthony J. Ries, and Kaleb McDowell

Human Research and Engineering Directorate, US Army Research Laboratory,
Aberdeen Proving Ground, MD 21005, USA
amar.marathe@case.edu, anthony.j.ries2@mail.mil,
kgm8@cornell.edu

Abstract. Machine learning techniques have been used to classify patterns of neural data obtained from electroencephalography (EEG) to increase human-system performance. This classification approach works well in controlled laboratory settings since many of the machine learning techniques used often rely on consistent neural responses and behavioral performance over time. Moving to more dynamic, unconstrained environments, however, introduces temporal variability in the neural response resulting in sub-optimal classification performance. This study describes a novel classification method that accounts for temporal variability in the neural response to increase classification performance. Specifically, using sliding windows in hierarchical discriminant component analysis (HDCA), we demonstrate a decrease in classification error by over 50% when compared to other state-of-the-art classification methods.

Keywords: Brain-Computer Interface (BCI), Rapid Serial Visual Presentation (RSVP), Electroencephalography (EEG), HDCA, Sliding HDCA, Temporal Variability, Single-trial, Real-world environment.

1 Introduction

Systems incorporating neural activity using EEG typically use machine learning techniques to classify or predict the occurrence of an action or event. To be useful, these systems must be able to function outside of the controlled confines of a laboratory setting. Moving into more dynamic environments introduces changes in the processing demands of the user as well as uncontrolled variability into the system. Variability of the EEG signal is influenced by an interaction of endogenous processes related to a user's state (e.g. fatigue), exogenous factors related to stimulus properties [1–4], and other system related factors. For example, it has been shown that the latency of the P300 event related potential (ERP) brain response is correlated with stimulus evaluation and reaction time [5, 6]. Stimuli that are easier to categorize produce faster reaction times and earlier P300 peak latencies than those that are more difficult to categorize. Thus, situations where the difficulty of stimulus categorization varies from trial to trial will produce a temporally variable neural response. Optimal performance

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 345–352, 2013.

of systems interpreting neural data must account for the existence of trial by trial temporal variability in the neural response.

Existing methods for single-trial classification can be divided into several categories. Some algorithms operate directly on the multi-channel EEG signals [7–11], while others apply spatial filters to transform the multi-channel EEG signal into a new signal that contains more task-relevant information prior to applying a standard machine-learning classifier [12–21]. Each of these existing methods have been shown to perform well in a specific task; however none of the previous studies has focused on testing the effects of temporal variability on classification performance. In this study, participants performed a rapid serial visual presentation (RSVP) target detection task. ERP analysis shows that the neural data contains large amounts of temporal variability. We show that a novel classification method that accounts for temporal variability can reduce classification error by over 50%.

2 Methods

2.1 Participants

Fifteen participants (9 male, age range 18-57, average age 39.5) volunteered for the current study. Participants provided written informed consent, reported normal or corrected-to-normal vision and reported no history of neurological problems. Fourteen of the fifteen participants were right-handed.

The voluntary, fully informed consent of the persons used in this research was obtained as required by Title 32, Part 219 of the Code of Federal Regulations and Army Regulations 70-25. The investigator has adhered to the policies for the protection of human subjects as prescribed in AR 70-25.

2.2 Stimuli and Procedure

Short video clips were used in a rapid serial visual presentation (RSVP) paradigm [22, 23]. Video clips either contained people or vehicles on background scenes, or only background scenes. Observers were instructed to make a manual button press with their dominant hand when they detected a person or vehicle (targets), and to abstain from responding when a background scene (distractor) was presented. Video clips consisted of five consecutive images each 100ms in duration; each video clip was presented for 500ms. There was no interval between videos such that the first frame was presented immediately after the last frame of the prior video. If a target appeared in the video clip, it was present on each 100ms image. The distracter to target ratio was 90/10. RSVP sequences were presented in two minute blocks after which time participants were given a short break. Participants completed a total of 25 blocks.

2.3 EEG Recording and Analysis

Electrophysiological recordings were digitally sampled at 512Hz from 64 scalp electrodes arranged in a 10-10 montage using a BioSemi Active Two system (Amsterdam,

Netherlands). External leads were placed on the outer canthus and below the orbital fossa of both eyes to record electrooculography (EOG). Continuous EEG data were referenced offline to the average of the left and right earlobes and digitally filtered 0.1-55Hz. To reduce muscle and ocular artifacts in the EEG signal and potential contamination with brain-based signals, we removed EOG and EMG artifacts using independent component analysis (ICA) [24].

ERP Analysis

ERP analysis was used to evaluate the trial by trial temporal variability of the neural response. Analyses for these data were previously reported [22] and are briefly described here. EEG data were processed and analyzed using EEGLAB [25] and ERPLab [26]. Continuous, artifact free data were epoched -1500 to 1500ms around target onset. Target epochs followed by a button press within 200 to1000ms and non-target epochs not followed by a response were included in the analysis. Averaging across all trials in a given condition may mask meaningful brain dynamics associated with performance; especially in perceptually difficult tasks in which the variance in ERP latency and reaction time (RT) increases [27]. Therefore, to assess the brain dynamics associated with varying levels of RT performance, target epochs were sorted into bins corresponding to an individual participant's reaction time quartile [28]. Grand averages across all subjects were then calculated for each quartile.

Single Trial Classification

The novel classification approach presented here is a modification of hierarchical discriminant component analysis (HDCA). Because of this, HDCA served as an ideal baseline measure of classification performance for this study. Details of the HDCA algorithm can be found in [7, 9–11] and it is briefly described below.

For classification purposes, EEG data were epoched -500 to 1600 ms around stimulus onset. Epoched EEG data were baseline corrected by removing the average of activity occurring between -500 and stimulus onset. Target epochs followed by a button press within 200 to 1000ms and all non-target epochs were included in the classification analysis.

Hierarchical Discriminant Components Analysis

HDCA transforms multi-channel EEG data collected over a temporal window relative to image onset into a single interest-score. Ideally, the interest score is generated so that the range of scores for each class are distinct, thereby allowing for simple discrimination of the two classes.

Generating interest scores from HDCA involves a two stage classification. In the first stage, our implementation uses a set of 15 discriminators applied to 15 non-overlapping 100 ms time windows that span 100 ms to 1600 ms after image onset. Each of the 15 discriminators is trained independently. Each discriminator combines the information contained in all 64 EEG signals collected over the course of the corresponding time window into a single value for discriminating target versus non-target. Thus, stage 1 of HDCA produces 15 interest scores that independently discriminate target from non targets. In the second stage, a separate discriminator is applied

to the output of the stage 1 discriminators to create a single interest score that can efficiently discriminate between target and non-target trials.

Sliding Hierarchical Discriminant Components Analysis
Sliding HDCA (sHDCA) builds upon the standard HDCA algorithm in an attempt to extract more information from temporally scattered events. sHDCA starts by using a standard HDCA classifier trained to discriminate targets versus non targets based on 500 ms of data between 300 ms and 800 ms after stimulus onset using 50 ms time slices. Rather than simply statically applying this classifier to each epoch, in sHDCA this initial classifier slides in time such that it is applied at each sample ranging from 200 ms prior to stimulus onset to 800 ms after stimulus onset. This sliding step means that the classifier is using epoch data from 100 ms post stimulus to 1600 ms post stimulus, which matches the data used by the standard HDCA algorithm.

Because each application of the standard HDCA algorithm produces a single score, sliding the HDCA classifier in time produces a single score per application (per time point). When the sliding process is complete, we are left with a score signal that is 1000 ms in duration. From this score signal, a second HDCA classifier is trained to discriminate targets versus non-targets based on the score signal. This second level classifier uses ten 100 ms time slices. The result of this HDCA classifier is the final score assigned to the epoch which is used to decide whether the current epoch is a target or non-target.

Cross Validation
A 10-fold cross validation was used to determine the accuracy for both classification methods. Data from each subject were divided into 10 equal sized blocks of trials. Classifiers were trained on 9 of the 10 blocks, and then tested on the block left out. This process was repeated 10 times such that each of the 10 blocks of trials was used as the independent testing set once. Performance was evaluated based on the area under the ROC curve (AUC). Each participant's performance was calculated as the average AUC calculated across all 10 cross validation sets. Statistical analyses for each classification method were performed on the average AUC for each participant.

Computational Requirements
Timing measures were also employed to evaluate the computational costs of training and testing each algorithm. For this evaluation, the MATLAB functions 'tic' and 'toc' were used to measure the total time needed for classifier training and testing. The time needed for testing was divided by the total number of trials in the test set to calculate an approximation of the total time needed to apply the classifier to a single epoch as would be required in a real-time application.

3 Results

3.1 Existence of Temporal Variability

Reaction time quartiles were used as binning parameters for the ERP analysis[28]. P3 latency exhibited a large amount of temporal variability relative to the stimulus onset

(Figure 1). P3 latency data were submitted to a one-way ANOVA with the main factor of Quartile containing four levels. Analysis showed a significant main effect of Quartile, $F(3,42) = 69.37$, $p < .001$. Subsequent t-tests revealed each quartile was significantly different ($\alpha = .05$) from each other after correction for multiple comparisons using Tukey's method indicating that P3 latency increased as RT became slower. (Figure 1).

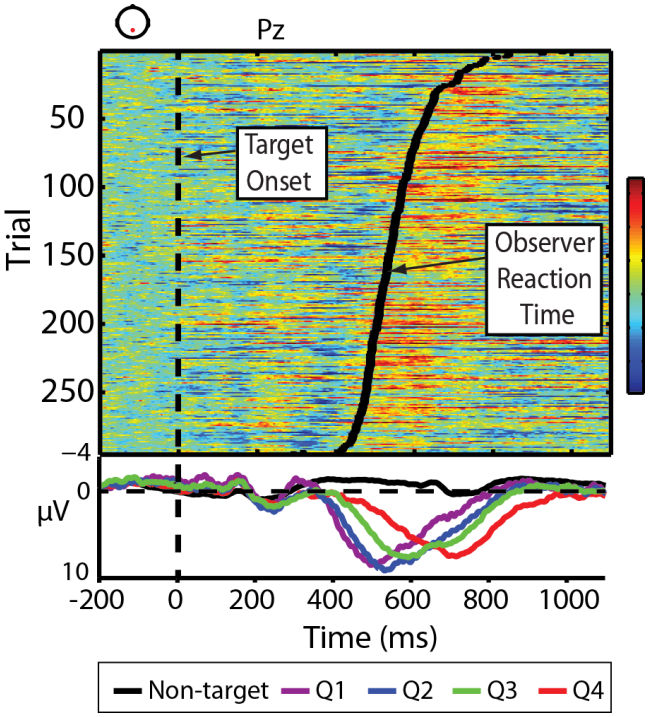


Fig. 1. Temporal variability in EEG of a single participant (S10). Upper plot shows single trial EEG response at Pz when activity is aligned to the target onset and sorted by response time. Lower plot shows average ERPs when reaction time is used as a binning parameter for ERP analysis.

3.2 Classification in the Face of Temporal Variability

Figure 1 clearly establishes the presence of temporal variability in the neural response. Figure 2 shows the accuracy of single-trial classification on these data. HDCA achieves a classification accuracy of 0.8691 ± 0.0359 (Mean AUC ± Std), while the classification accuracy of Sliding HDCA was 0.9365 ± 0.0223 (mean ± std AUC). This represents a 51.5% reduction of classification error and the overall difference is statistically different (Wilcoxon Sign Rank Test $p<0.001$).

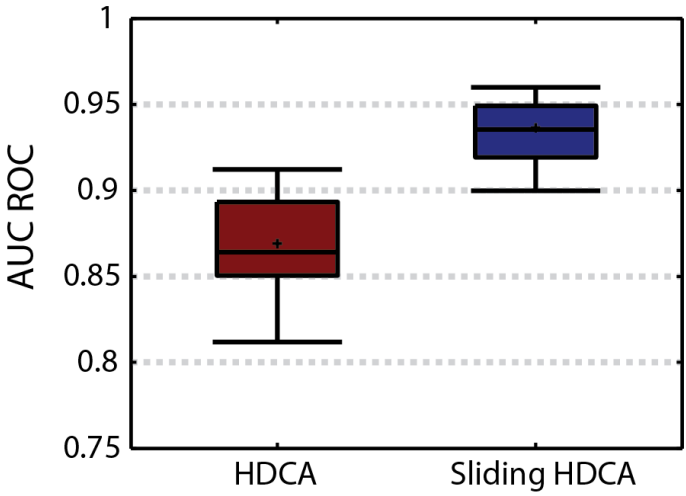


Fig. 2. Classification results across 15 subjects. Horizontal lines in each box represent the median and the dot represents the mean. The maroon box shows the classification accuracy when using the standard HDCA algorithm. The blue box shows the classification accuracy when using sliding HDCA. The difference is significant using the Wilcoxon Sign Rank Test ($p < 0.001$).

3.3 Execution Time

Sliding HDCA represents a potential improvement upon the standard HDCA classification scheme but comes at the cost of increased computing time. Training a standard HDCA classifier on the data set described here takes approximately 10 to 15 seconds. Training a sliding HDCA classifier on the same data set using the parameters described above takes 354 ± 33 seconds – a 20 to 35 fold increase in training time. Applying a standard HDCA classifier to this data set typically takes less than a millisecond per epoch, while applying sliding HDCA takes 383 ± 4 ms. While these relative time comparisons are important, in most RSVP applications, requiring approximately 6 minutes to train a classifier and 383 ms to apply the classifier is perfectly reasonable.

4 Discussion

The current study employed a dynamic RSVP task using short-duration videos. ERP analyses showed a high degree of temporal variability in the neural response. This study developed a novel classification scheme that overcame the temporal variability in the data without needing to use information from the behavioral response.

Sliding HDCA classification is a novel classification method described here that reduced classification error by over 50% over a standard HDCA classifier using the same amount of data. The increased accuracy of sHDCA classification comes at the expense of computation time. The increase in computation time is significant; however for most applications the increased accuracy seen with sHDCA will far outweigh the increase in computation time.

This study demonstrates that algorithms that account for temporal variability can dramatically improve classification accuracy. The novel method described here is one such method. This method enables further development of applications that either replace or augment behavioral responses for tasks where variable reaction times are expected.

5 Conclusion

The Sliding HDCA method described here provides a means to overcome the temporal variability in the neural response that is likely to occur in more complex environments. By transforming the raw EEG signal into a score signal, the sliding step of sHDCA produces a new signal that emphasizes the discriminating features of the EEG input and consequently improves single trial classification. The efficacy of this approach was demonstrated in an RSVP target detection task; however this approach may also prove to be useful for other types of BCI technologies in which temporal variability causes a drop in performance.

References

1. Kammer, T., Lehr, L., Kirschfeld, K.: Cortical visual processing is temporally dispersed by luminance in human subjects. Neuroscience Letters 263, 133–136 (1999)
2. Folstein, J.R., Van Petten, C.: After the P3: Late executive processes in stimulus categorization. Psychophysiology 48, 825–841 (2011)
3. Craig, A., Tran, Y., Wijesuriya, N., Nguyen, H.: Regional brain wave activity changes associated with fatigue. Psychophysiology 49, 574–582 (2012)
4. Lal, S.K.L., Craig, A.: Driver fatigue: Electroencephalography and psychological assessment. Psychophysiology 39, 313–321 (2002)
5. Magliero, A., Bashore, T.R., Coles, M.G.H., Donchin, E.: On the Dependence of P300 Latency on Stimulus Evaluation Processes. Psychophysiology 21, 171–186 (1984)
6. Dien, J., Spencer, K.M., Donchin, E.: Parsing the late positive complex: Mental chronometry and the ERP components that inhabit the neighborhood of the P300. Psychophysiology 41, 665–678 (2004)
7. Gerson, A.D., Parra, L.C., Sajda, P.: Cortically coupled computer vision for rapid image search. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14, 174–179 (2006)
8. Tomioka, R., Aihara, K., Müller, K.R.: Logistic regression for single trial EEG classification. Advances in Neural Information Processing Systems 19, 1377–1384 (2007)
9. Parra, L.C., Christoforou, C., Gerson, A.D., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M.G., Sajda, P.: Spatiotemporal Linear Decoding of Brain State. IEEE Signal Processing Magazine 25, 107–115 (2008)
10. Sajda, P., Pohlmeyer, E., Wang, J., Parra, L.C., Christoforou, C., Dmochowski, J., Hanna, B., Bahlmann, C., Singh, M.K., Chang, S.-F.: In a Blink of an Eye and a Switch of a Transistor: Cortically Coupled Computer Vision. Proceedings of the IEEE 98, 462–478 (2010)
11. Pohlmeyer, E.A., Wang, J., Jangraw, D.C., Lou, B., Chang, S.-F., Sajda, P.: Closing the loop in cortically-coupled computer vision: a brain–computer interface for searching image databases. Journal of Neural Engineering 8, 036025 (2011)

12. Ramoser, H., Muller-Gerking, J., Pfurtscheller, G.: Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Transactions on Rehabilitation Engineering 8, 441–446 (2000)
13. Lemm, S., Blankertz, B., Curio, G., Muller, K.-R.: Spatio-spectral filters for improving the classification of single trial EEG. IEEE Transactions on Biomedical Engineering 52, 1541–1548 (2005)
14. Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., Muller, K.-R.: Combined Optimization of Spatial and Temporal Filters for Improving Brain-Computer Interfacing. IEEE Transactions on Biomedical Engineering 53, 2274–2281 (2006)
15. Tomioka, R., Dornhege, G., Nolte, G., Blankertz, B., Aihara, K., Müller, K.R.: Spectrally weighted common spatial pattern algorithm for single trial EEG classification. Dept. Math. Eng., Univ. Tokyo, Tokyo, Japan, Tech. Rep. 40 (2006)
16. Wu, W., Gao, X., Hong, B., Gao, S.: Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL). IEEE Trans. Biomed. Eng. 55, 1733–1743 (2008)
17. Farquhar, J.: A linear feature space for simultaneous learning of spatio-spectral filters in BCI. Neural Networks 22, 1278–1285 (2009)
18. Cecotti, H., Rivet, B., Congedo, M., Jutten, C., Bertrand, O., Maby, E., Mattout, J.: A robust sensor-selection method for P300 brain–computer interfaces. Journal of Neural Engineering 8, 016001 (2011)
19. Touryan, J., Gibson, L., Horne, J.H., Weber, P.: Real-Time Measurement of Face Recognition in Rapid Serial Visual Presentation. Front Psychol. 2 (2011)
20. Yu, K., Shen, K., Shao, S., Ng, W.C., Kwok, K., Li, X.: Common Spatio-Temporal Pattern for Single-Trial Detection of Event-Related Potential in Rapid Serial Visual Presentation Triage. IEEE Transactions on Biomedical Engineering 58, 2513–2520 (2011)
21. Yu, K., Shen, K., Shao, S., Ng, W.C., Li, X.: Bilinear common spatial pattern for single-trial ERP-based rapid serial visual presentation triage. Journal of Neural Engineering 9, 046013 (2012)
22. Ries, A.J., Larkin, G.B.: Stimulus and response-locked P3 activity in a dynamic RSVP task. ARL-TR-6314 (2012)
23. Touryan, J., Gibson, L., Horne, J.H., Weber, P.: Real-time classification of neural signals corresponding to the detection of targets in video imagery. Presented at the International Conference on Applied Human Factors and Ergonomics, Miami, Fl (2010)
24. Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M.J., Iragui, V., Sejnowski, T.J.: Removing electroencephalographic artifacts by blind source separation. Psychophysiology 37, 163–178 (2000)
25. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134, 9–21 (2004)
26. Luck, S.J., Lopez-Calderon, J.: ERPLAB Toolbox (2010)
27. Luck, S.J.: An Introduction to the Event-Related Potential Technique. A Bradford Book (2005)
28. Poli, R., Cinel, C., Citi, L., Sepulveda, F.: Reaction-time binning: A simple method for increasing the resolving power of ERP averages. Psychophysiology 47, 467–485 (2010)

A Translational Approach to Neurotechnology Development

Kaleb McDowell and Anthony Ries

Human Research and Engineering Directorate, U.S. Army Research Laboratory,
Aberdeen Proving Ground, MD 21005, USA
Kgm8@cornell.edu, anthony.j.ries2.civ@mail.mil

Abstract. The past several decades have seen an explosion of meaningful and nuanced insights into the connection between human behavior and the nervous system; however, the translation of these insights into viable applications is a non-trivial and widely acknowledged challenge. Recent advancements in brain-computer interaction and real-world neuroimaging technologies have provided major breakthroughs that provide the underpinnings for translational neuroscience research efforts. This session focuses on building off of those advancements and specifically proposes three concepts necessary for overcoming the challenges of translation: 1) integrating aspects of knowledge of brain function that are generally separate into single analyses, 2) increasing situational complexity, and 3) continuing to develop neuroimaging tools specifically for use in real-world environments.

Keywords: Translational Neuroscience, Neurotechnology, Brain-Computer Interface (BCI), Electroencephalography (EEG), Neural Classification.

1 Introduction

The past several decades have seen an explosion of meaningful and nuanced insights into the connection between human behavior and the nervous system; a connection that is considered to be foundational for understanding how we perceive and interact with the external world. Underlying the discovery of these insights is the highly active field of neuroscience, which has generated approximately ½ million citable documents over the past 15 years [1]. Analysis and recommendations from numerous sources including the Office of the U.S. President and the National Research Council suggest that the continued advancements in neuroscience have the potential to revolutionize human-system integration technologies and have a dramatic impact on society at large [2–4]. However, translating basic neuroscience research into viable applications is a non-trivial and widely acknowledged challenge, particularly within the medical community [5].

The difficulties in translation are due, in large part, to the overwhelming complexity of the human brain and its approximately 100 billion neurons. Most basic neuroscience research has been largely conducted using a reductionist approach in which experimental

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 353–360, 2013.

variables are highly controlled and both stimuli and behavioral responses are relatively simple compared to those found in real-world settings. These highly constrained laboratory environments are often used to enhance signal-to-noise ratios, and limit the behaviors and interactions study participants normally have with the world. Although this research has advanced our basic understanding of brain function within highly constrained environments, the extent to which controlled laboratory research results generalize to brain function in complex and dynamic real-world environments is currently not well understood [6, 7]. Further, it has been argued that the complexity of the human brain allows not only different individuals to process information differently, but also the same individuals may engage different brain structures to cognitively process similar information, particularly in response to contextual changes [8]. These concepts suggest that there may be fundamental differences in real-world brain function relative to that observed in highly controlled laboratory environments, thus supporting the need for more ecological approaches [9] focusing on human, task, and environmental interactions [10] within real-world environments. We posit that such approaches would lead to a more representative understanding of real-world brain function and enhance the success of translational neuroscience efforts.

2 Translational Advances in Human-System Integration

The potential for translating neuroscience into revolutionary human-system integration technologies has been recently boosted by technological advancements in two areas: the rise of brain-computer interaction (BCI) technologies and the advent of real-world neuroimaging tools. These advancements have enhanced not only the ability to measure neural activity in real-world situations, but also the ability to interpret neural data generated within complex scenarios. As discussed below, the technologies and insights generated in these two fields are now enabling the ecological approaches needed for successful translational neuroscience.

The past decade had seen a surge of research and technological advancements in BCI and associated technologies, particularly as applied to the medical domain. Many of the advancements in BCI arose from the unique combinations of researchers focused on the man-machine problem as opposed to simply addressing the traditional neuroscience questions underlying nervous system function. Early BCI efforts focused on providing communications and direct control capabilities to specific clinical populations [11]. However, the mix of scientific and engineering approaches have changed the focus of the original BCI efforts (e.g., for discussion, see [12]), and provided novel insights into not only how neuroimaging could be effectively integrated in system designs, but also how the brain functions. These understandings are allowing researchers to uncover approaches to integrate emotion into video games, toys, advertising, and music [13, 14], merge human pattern recognition with computer processing power for joint human–computer object detection [15, 16], and overall, develop applications that use neural signals in ways that are more consistent with the brain function that naturally occurs during task performance [17]. Importantly, the efforts in this domain have augmented current methodologies and given rise to a wide variety of novel research approaches and tools for analyzing and interpreting neural

signals in complex settings. For example, the BCI community has made dramatic improvement in real-time signal processing and the use of machine learning approaches including artificial neural networks and signal classification methods for neuroimaging.

Driven by a focus on application, recent advances in BCI technologies have provided proof-of-principle that neurotechnology applications can be effective outside clinical and controlled laboratory settings; however, the state-of-the-art in this area is still limited in its utility. For example, direct control BCIs are becoming viable for clinical populations; but, their performance is still well below that of healthy users employing traditional optimized control devices such as a keyboard, mouse, or joystick. In the near future, we expect greater translation into the BCI field as the focus of BCI's shift towards use by healthy populations. We foresee three technical areas aiding the translational effort: more effective approaches to integrating neural processing into BCI design, the integration of BCIs into more complex situations, and a shifting focus to BCIs that accomplish unique tasks that are difficult to accomplish through other means (for examples, see [4, 18]).

The second major advancement over the past decade has been the development of neuroimaging tools for use in operationally relevant settings, which has been enabled in large part due to DoD programs such as Augmented Cognition [19]. Neural sensing technology is one area that has seen several significant advancements in recent years (for a full review, see [20]). Specifically, electroencephalography (EEG) has shown the most promise as a near-term solution to the challenges of quality, mobility and wearability within realistic operational environments. Recent advancements occurring in the areas of dry, comfortable EEG electrodes and wireless EEG systems have shown particular promise. A second area of advancement has been in EEG analytic techniques and approaches (for review, see [21]) and accessible software for both off-line and real-time EEG analysis. These tools are enabling steady progress toward real-world neuroimaging capabilities. Significant conceptual progress in both hardware and software still needs to be accomplished to enable end state goals of wear-and-forget sensing technologies capable of producing laboratory grade results in real-time and in environmental conditions never before deemed possible.

3 Three Translational Concepts for Human-System Integration

In this session, we highlight three concepts that build off of the advancements in BCI and real-world neuroimaging technologies to enable the translation of neuroscience needed to provide revolutionary advances in military neurotechnologies projected by the National Research Council and others (e.g., [3]). The first three talks of the session focus on improving signal analysis, classification, and interpretation through a integrating a better understanding of nervous system function into classifier design. In the fourth talk, the speaker will focus on the translation of a BCI into operationally complex situation. The final talks of the session will focus on tools for improving real-world neuroimaging and neurotechnologies.

3.1 Integrating Multiple Aspects of Neuroscience Knowledge into Single Analyses

The vast extant neuroscience literature is filled with important and interesting insights into brain function. However, many of the neuroscience studies and applications to date focus on isolated research areas, avoiding potential confounds or contextual modulations that make interpretation difficult. For example, much of the direct control BCI literature avoids the psychological construct of mental fatigue by designing studies that are limited in duration or by throwing out data when subjects seem fatigued. In the first part of this session, we focus on attempting to merge different conceptualizations of nervous system function into single analyses to overcome some of the aforementioned confounds or contextual modulations that make interpretation difficult. Specifically, in the first three talks, the speakers focus on combining known but underutilized information about brain function into more traditional analyses for improved performance:

- Amar Marathe and colleagues examine the classification of targets from neural data as participants perform a rapid, serial visual presentation (RSVP) task [22]. According to the literature, during complex situations, temporal variability exists between stimulus presentation and the measurement of neural responses due to factors both internal and external to the operator. However, to date, target classifier schemes have not effectively accounted for this trial-by-trial variability. Dr. Marathe presents a study that describes a novel classification method that accounts for temporal variability in the neural response to increase classification performance and behavioral prediction. Specifically, using sliding windows in hierarchical discriminant component analysis (HDCA), they demonstrate a decrease in classification error by over 50% when compared to a state-of-the-art HDCA method.
- Jon Touryan and colleagues examine the generalizability of fatigue-based measures of EEG to predict task performance [23]. As previously mentioned, BCI research has often avoided addressing fatigue.. Dr. Touryan presents a study that takes the first steps towards developing an RSVP-based BCI that continues to effectively function as operators become fatigued. He presents a study that extends the fatigue-based performance prediction algorithms developed for the driving domain [24] to RSVP performance prediction and demonstrate similar results for both tasks. This study illustrates the capability to detect the state-based information stream within an existing BCI task (i.e., RSVP) that is needed to extend BCI algorithms that adapt to user state.
- Greg Apker and colleagues examine the ability to use fatigue-based measures of EEG to predict driving performance [25]. Previous research has shown the ability to make such predictions based on findings of a linear relationships between power spectral density estimates of EEG and driving performance in simple driving tasks [24] However, the extant literature also suggests that this simple relationship is insufficient based on multiple findings including: 1) task performance depends on numerous factors in addition to fatigue and 2) state changes in the brain may produce non-linearities in the relationship between the EEG signal and behavior.

Dr. Apker presents a study that merges a quadratic discriminate classifier with a modified version of the original linear approach and demonstrates prediction improvements through the inclusion of the non-linear element.

3.2 Increasing Situational Complexity

Theoretical and experimental evidence suggests that there are fundamental differences in how the human brain functions to control behavior when it is situated in ecologically-valid environments (i.e., situated cognition) versus that observed in highly-controlled laboratory environments. We hypothesize that a portion of our understanding of brain function will generalize to real-world settings, and from a translational perspective, it is critical to identify which portion. This generalizability issue defines the need to expand the capability of neuroimaging technologies beyond the laboratory and into complex situations where natural human, task, and environmental interactions can be studied. The fourth talk of the session illustrates the successful translation of a generalizable research finding to a complex scenario:

- Anthony Ries presents data supporting the use of the RSVP task for searching imulated urban environments with performance improvements over a manual search of the same environment. Second, Dr. Ries presents the development underlying a novel simulation environment designed to aid the translation of the RSVP into a more realistic context. Specifically, the novel simulation environment embeds the RSVP into an operationally-relevant multitasking scenario where the operator is required to search for targets, identify IEDs near the roadside, and respond to specific radio communications while riding in a simulated moving vehicle. The initial results of a validation study support the successful translation of the RSVP task into the more complex environment with the neural-driven target detection approach outperforming manual target detection [26].

3.3 Developing Neuroimaging Tools Specifically for the Real-World

The capability to extend both science and applications into a wider variety of real-world situations will be critical to the effectiveness of translational neuroscience. Developing neuroimaging tools and in particular tools that function in a wide range of settings is a crucial component of such a capability and is a top research priority according to the Executive Office of the President [2]. In the final part of this session, two speakers will present improved tools for collecting and interpreting neural signals in real-world situations:

- Bret Kellihan and colleagues describe a neuroimaging tool for understanding the human brain's interaction with real-world stress [27]. While laboratory settings have offered a great deal of insight into the brain function underlying stress, the effectiveness of laboratory stressors to represent the entire span of real-world stressors has been called into question [28]. Mr. Kellihan presents a paper discussing the state-of-the-art in real-world stress measurement technologies and the limitations of current systems. He also describes the novel multi-aspect real-world

neuroimaging system (MARIN) that has been developed specifically for studying the brain under conditions of real-world stress. This tool will enable a better understanding of the neurological mechanisms of stress and enable the development of future neurotechnologies that function in real-world situations.

- Vernon Lawhern and colleagues focus on the critical issue of artifact or non-brain electrical signal identification within EEG time series [29]. Real-world environments pose critical issues with non-brain electrical signals that are dramatically different than the issues posed under laboratory conditions designed to minimize sources of artifact. Hence, real-world neuroimaging will require effective tools to eliminate or potentially separate and utilize the non-brain from brain electrical sources. Dr. Lawhern presents a study that focuses on identifying artifacts using autoregressive (AR) models and specifically proposes a method for determining optimal AR features based on a penalized multinomial regression. The authors' results indicate that the size of the feature vector can be greatly reduced with minimal loss to classification accuracy, which has significant ramifications for both computation efficiency and hardware design.

Together, these talks on neurotechnology development highlight translational advancements in BCI and real-world neuroimaging technologies. The multifaceted approach taken by the authors in this session demonstrates the importance of developing neurotechnology tools and methods to enhance human system performance as well as a means to measure the brain-in-action within ecologically valid environments.

References

1. Scimago Journal & Country Rank, http://www.scimagojr.com/
2. Social, behavioral, and economic research in the federal context: report of the National Science and Technology Council Subcommittee on Social, Behavioral, and Economic Sciences (January 2009)
3. Committee on Opportunities in Neuroscience for Future Army Applications; National Research Council: Opportunities in Neuroscience for Future Army Applications. The National Academies Press, Washington, D.C (2009)
4. Lance, B.J., Kerick, S.E., Ries, A.J., Oie, K.S., McDowell, K.: Brain-Computer Interface Technologies in the Coming Decades. Proceedings of the IEEE 100, 1585–1599 (2012)
5. Academic Health Centers: Leading Change in the 21st Century, http://www.nap.edu/openbook.php?isbn=0309088933
6. Kingstone, A., Smilek, D., Ristic, J., Friesen, C.K., Eastwood, J.D.: Attention, Researchers! It Is Time to Take a Look at the Real World. Current Directions in Psychological Science 12, 176–180 (2003)
7. Kerick, S.E., McDowell, K.: Understanding brain, cognition, and behavior in complex dynamic environments. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) Augmented Cognition, HCII 2009. LNCS, vol. 5638, pp. 35–41. Springer, Heidelberg (2009)
8. Edelman, G.M., Gally, J.A.: Degeneracy and complexity in biological systems. PNAS 98, 13763–13768 (2001)
9. Gibson, J.J.: The theory of affordances. In: Perceiving, Acting, and Knowing. Towards an Ecological Psychology. John Wiley & Sons Inc., Hoboken (1977)

10. Gevins, A., Leong, H., Du, R., Smith, M.E., Le, J., DuRousseau, D., Zhang, J., Libove, J.: Towards measurement of brain function in operational environments. Biological Psychology 40, 169–186 (1995)
11. Wolpaw, J., Birbaumer, N., McFarland, D., Pfurtscheller, G., Vaughan, T.: Brain-computer interfaces for communication and control. Clinical Neurophysiology 113, 767–791 (2002)
12. Lance, B.J., Kerick, S.E., Ries, A.J., Oie, K.S., McDowell, K.: Brain Computer Interface Technologies in the Coming Decades. arXiv:1211.0886 (2012)
13. Nijholt, A., Tan, D.: Playing with your brain: brain-computer interfaces and games. In: Proceedings of the International Conference on Advances in Computer Entertainment Technology, pp. 305–306. ACM Press, New York (2007)
14. Show Your Mood with Brain-Controlled "Necomimi" Cat Ears - Neurogadget.com, http://neurogadget.com/2011/05/06/show-your-mood-with-brain-controlled-necomimi-cat-ears/2100
15. Sajda, P., Pohlmeyer, E., Wang, J., Parra, L.C., Christoforou, C., Dmochowski, J., Hanna, B., Bahlmann, C., Singh, M.K., Chang, S.-F.: In a Blink of an Eye and a Switch of a Transistor: Cortically Coupled Computer Vision. Proceedings of the IEEE 98, 462–478 (2010)
16. Poolman, P., Frank, R.M., Luu, P., Pederson, S.M., Tucker, D.M.: A single-trial analytic framework for EEG analysis and its application to target detection and classification. Neuroimage 42, 787–798 (2008)
17. Wolpaw, J.R.: Brain–computer interfaces as new brain output pathways. J. Physiol. 579, 613–619 (2007)
18. Zander, T.O., Kothe, C.: Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. Journal of Neural Engineering 8, 025005 (2011)
19. Schmorrow, D.: Foundations of Augmented Cognition. Erlbaum (2005)
20. Liao, L.-D., Lin, C.-T., McDowell, K., Wickenden, A.E., Gramann, K., Jung, T.-P., Ko, L.-W., Chang, J.-Y.: Biosensor Technologies for Augmented Brain-Computer Interfaces in the Next Decades. Proceedings of the IEEE 100, 1553–1566 (2012)
21. Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo, N., Zhang, Z., Kreutz-Delgado, K.: Evolving Signal Processing for Brain-Computer Interfaces. Proceedings of the IEEE 100, 1567–1584 (2012)
22. Marathe, A., Ries, A.J., McDowell, K.: A novel method for single-trial classification in the face of temporal variability. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2013. LNCS (LNAI), vol. 8027, pp. 345–352. Springer, Heidelberg (2013)
23. Touryan, J., Apker, G., Kerick, S., Lance, B., Ries, A.J., McDowell, K.: Translation of EEG-based performance prediction models to rapid serial visual presentation tasks. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2013. LNCS (LNAI), vol. 8027, pp. 521–530. Springer, Heidelberg (2013)
24. Lin, C.-T., Wu, R.-C., Jung, T.-P., Liang, S.-F., Huang, T.-Y.: Estimating Driving Performance Based on EEG Spectrum Analysis. EURASIP Journal on Advances in Signal Processing 2005, 521368 (2005)
25. Apker, G., Lance, B., Kerick, S., McDowell, K.: Combined linear regression and quadratic classification approach for an EEG-based prediction of driver performance. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2013. LNCS (LNAI), vol. 8207, pp. 231–240. Springer, Heidelberg (2013)

26. Touryan, J., Ries, A.J., Weber, P., Gibson, L.: Integration of automated neural processing into an army-relevant multitasking simulation environment. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2013. LNCS (LNAI), vol. 8027, pp. 790–798. Springer, Heidelberg (2013)
27. Kellihan, B., Doty, T.J., Hairston, W.D., Canady, J., Whitaker, K.W., Lin, C.-T., Jung, T.-P., McDowell, K.: A real-world neuroimaging system to evaluate stress. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2013. LNCS (LNAI), vol. 8027, pp. 316–325. Springer, Heidelberg (2013)
28. Zanstra, Y.J., Johnston, D.W.: Cardiovascular reactivity in real life settings: measurement, mechanisms and meaning. Biol. Psychol. 86, 98–105 (2011)
29. Lawhern, V., Hairston, W.D., Robbins, K.: Optimal feature selection for artifact classification in EEG time series. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2013. LNCS (LNAI), vol. 8027, pp. 326–334. Springer, Heidelberg (2013)

Understanding Brain Connectivity Patterns during Motor Performance under Social-Evaluative Competitive Pressure

Hyuk Oh[1,2,*], Rodolphe J. Gentili[1,2,3], Michelle E. Costanzo[1,2,4], Ronald N. Goodman[1,2,5], Li-Chuan Lo[2], Jeremy C. Rietschel[1,2,5], Mark Saffer[1,2], and Bradley D. Hatfield[1,2]

[1] Neuroscience and Cognitive Science Program
[2] Department of Kinesiology
[3] Maryland Robotics Center, University of Maryland
College Park, MD 20742, USA
[4] Now with the Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA
[5] Now with the Baltimore VA Medical Center, Baltimore, MD 21201, USA
hyukoh@umd.edu

Abstract. Previous studies have shown that psychological arousal impacts motor performance during social-evaluative tasks by its influence on cortical dynamics, which can translate into motor performance enhancement. Although these findings have established critical links between performance under mental stress and elevated brain activity beyond that required for performance, there is still a need to further investigate brain connectivity during cognitive motor performance under such conditions. Here both electroencephalographic (EEG) and shooting performance were obtained in a shooting task under both performance-alone and competitive conditions. Network connectivity was assessed for the localized EEG sources. The results are consistent with those previously obtained and suggest elevated statistical dependencies and causal interactions between motor and non-motor areas during the competitive condition relative to performance-alone. Such network analysis provides a complementary approach to more traditional EEG derived metrics allowing for examining brain dynamics during cognitive motor performance under varying conditions of mental stress.

Keywords: Brain connectivity, EEG Localization, Motor Cognition, Competitive Pressure.

1 Introduction

Some individuals are better able to perform under high pressure, while others fail to perform up to their skill and ability (i.e., choking under pressure [1]). For example,

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 361–370, 2013.

social-evaluative stress such as competition often leads to significant fluctuations in the quality of motor performance [2]. In such situations, it has been reported that not only physiological factors (e.g., circulatory and electromyography) but also neurocognitive aspects (e.g., mental state and neural processes) play a critical role in the quality of motor performance [3–6]. Several studies have demonstrated that experts employ less verbal-analytical processing during skilled motor performance, resulting in attenuation of nonessential cognitive motor processes possibly due to a shift to reliance on subcortical structures and relative engagement of visuospatial processing [7]. A recent multilevel examination of motor performance and cortical dynamics under social-evaluative competitive pressure [8] found a loss of psychomotor efficiency during competitive performance; i.e. elevation of non-essential neural activity and cerebral cortical networking. Namely, during competition relative to a non-competitive (i.e., performance-alone) condition, each measure was respectively reported as dysfluency of the aiming trajectory, modestly elevated physiological responses, and increased cortico-cortical communication between motor and other brain regions, accompanied by relative desynchrony of high alpha power [8].

Such performance changes may be a consequence of reinvestment, in which a performer focuses explicit attention and control to well learned motor skills during mental stress exposure, which, in turn, results in performance degradation [9]. Thus, the confluence of increased state anxiety and explicit self monitoring leads to conscious control of essential motor control processes such that the performer reverts from the advanced stage of automaticity to an earlier stage of effortful analysis where verbal-analytical processing interferes with the refinement of skilled action. In addition social evaluative pressure may also act to increase the cognitive-motor task difficulty or workload load during performance resulting in elevated neural effort during task execution [10]. Thus, it seems to be reasonable that increased cortical activation (especially in verbal temporal regions) could occur during social-evaluative competitive pressure and could disrupt psychomotor efficiency [5].

To better understand how performance under competitive pressure relates to elevated neural activity beyond that required for performance, this study uses a novel EEG tomography techniques called low resolution brain electromagnetic tomography algorithm (LORETA) to identify the three dimensional (3D) distribution of the generating electric neuronal activity [11]. In LORETA, the source space is restricted to gray matter and the hippocampus as determined in the digitized probability atlas based on the Talairach human brain atlas (Brain Imaging Center, Montreal Neurological Institute (MNI)). Based on response similarity in such localized EEG sources, connectivity (e.g., structural, functional, or effective connectivity) analysis is processed to infer interregional communications between several brain regions. This study examined the EEG source distribution under competitive pressure as compared with non-competitive condition using LORETA, and to identify the brain connectivity from performers under varying conditions of mental stress by correlating cortical dynamics with motor performances.

2 Methods

2.1 Data Acquisition

Participants completed a dry fire (i.e., no ammunition) pistol shooting task under performance-alone (PA) and competitive (C) conditions while EEG and shooting performance were recorded.

Subjects. Nineteen subjects (17 men and 2 women; age range of 18-38 years; mean and standard deviation age of 22 and 4.33), enrolled in the Reserve Officers' Training Corps (ROTC) program, participated in the present study. All subjects were right-hand dominant and right-eye dominant and reported no history of neurological or psychiatric disorders as well as psychotropic medications at the time of their participation in the study. In addition, all subjects met a minimum performance level for inclusion in to the study such that each participant had to hit the target 80% of the time or greater during a preliminary practice session consisting of 40 shots. Prior to testing, all participants granted their written informed consent in accordance with the protocol approved by the University of Maryland Institutional Review Board, and were also informed that they were free to withdraw from the study at any time.

EEG Measures. EEG data were acquired from 30 EEG channels (Fig. 1b) in accordance with the 10-20 system using a linked earlobes reference and a common ground on FPz with 2 bipolar electrooculography (EOG) channels (horizontal HEOG at the outer canthi of both eyes and vertical VEOG placed above and below the left eye over the orbicularis oculi muscle). The data were recorded with an online bandpass filter at 0.01-100 Hz and a sampling rate of 1000 Hz using SCAN 4.3.3 (Compumedics Neuroscan, Charlotte, NC, USA). EEG baselines (1 min standing in shooting position without pistol) were collected prior to each session commencement.

Shooting Tasks. A dry fire pistol shooting task was completed in a sound attenuated testing chamber, for which a prism technique based shooter training system, Noptel ST-2000 version 2.33 (Noptel Oy, Oulu, Finland), was used to monitor the shooting performance at 66 Hz: e.g., both the position of the instantaneous aiming point and the shot placement in mm on the target as well as shot score. Participants shot from a standing position 5 m (Fig. 1a) from an appropriately scaled target to maintain a proportionate diameter consistent with that of a standard competitive target at a distance of 50 feet (or 15.24 m). They held the pistol with their dominant (i.e., right) hand and had their nondominant (i.e., left) eye occluded.

Two testing conditions (PA and C) were counterbalanced such that half of the participants engaged in PA followed by C, and the other half of them completed C first and then PA with a 15 min rest period in between to ensure a stable attention state and to minimize the adverse effects of fatigue. Participants were allowed 10

practice shots prior to each testing condition, and completed 40 self-paced shots (a 30 s time constraint for each shot in condition C) during both PA and C conditions. In each shot (or trial), an electronic pulse was generated by the Noptel to mark the trigger pull in the continuous EEG recording, and visual feedback on shot placement on the target as well as shot score was provided after each shot. The shot location was recorded as the position of the aiming point on the target at the time of the trigger pull, and shot score was proportional to the proximity of the hit point from the bullseye: a maximum of 10 points at the bullseye, and 1 point at least touching the outermost ring.

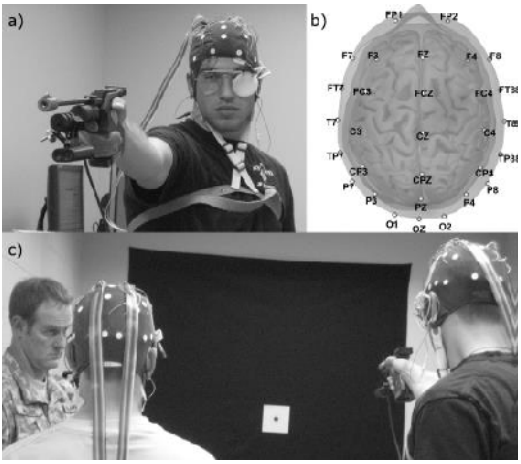


Fig. 1. (a), (c) A shooting performance was monitored using a prism technique based shooter training system during both PA and C conditions. During condition C, two participants took turns shooting at the target in a social evaluation environment by a superior officer and the opponent. b) 30 EEG electrode placements on the international 10-20 system. EOG, reference, and ground electrodes are not shown.

Condition PA. Participants were not evaluated but instructed to remain focused and relaxed during this condition. Following the baseline measures and the practice shots, the first 20 shots for record (i.e., block 1) were executed followed by a 5 min break, and then the final 20 shots for record (i.e., block 2) were executed (Fig. 1a).

Condition C. This condition involved the same order of measurements as PA, but included direct comparison of shooting performance to another study participant. Two participants took turns shooting at the target such that one shot while the other observed the opponent's performance and the shooting order was alternated across trials (Fig. 1c). Participants were instructed to set the pistol down between each shot and to remain standing throughout the respective conditions. Scores were presented to the competitor after each trial (i.e. after both participants had taken one shot) and a winner of that trial was declared. Participants were explicitly informed of all of the

following testing rules to exert competitive pressures prior to task execution and were encouraged to win the competition:

- social evaluation by a superior officer who conspicuously took notes and evaluated the participants' shooting stance and accuracy,
- financial loss or gain of 50 cents per round from a starting sum of $20; in the case of a tie, the sum at stake (i.e., $1) carried over to the next round; a dollar bonus or loss respectively for a bullseye or missing the target completely,
- a 30 s time constraint for each shot, beginning when the participant first grasped the pistol to initiate the shooting position,
- video camera recording,
- social responsibility as a team member; participants were placed on teams such that their score contributed to overall team score, both of which were displayed outside the ROTC field house.

2.2 Data Analysis

Preprocessed EEG data were localized by applying LORETA, and then statistical dependencies (e.g., cross-spectral connectivity) between localized EEG sources were investigated along with the co-registered shooting performance.

Shooting Data. The time for each shot started around -4 s before trigger pull, which corresponds to time zero. Participants performed motionless precision aiming tasks, allowing for minimal artifacts up to the time of the trigger pull. Aiming variability was quantified as the standard deviation of the tangential displacement of the shot placement with respect to the position of the aiming point at 3 s prior to trigger pull (instead of 4 s to minimize artifacts). In addition, mean shooting scores across shots for each subject on each condition were computed.

EEG Data. Ocular artifacts were reduced from the EEG data by employing a regression procedure with artifact averaging method using SCAN 4.3.3: a positive deflection with the trigger threshold of 10 % from the maximum artifact voltage, and 20 and 400 for minimum sweeps and the sweep duration in ms, respectively. The EEG data were then visually inspected to reject any trials that still contained significant artifacts, and were bandpass filtered by a bidirectional Butterworth filter between 3 to 50 Hz with a 24 dB/octave rolloff. Next, the continuous EEG data were partitioned to have a 3 s period of EEG data points prior to trigger pull (as shooting data) in each trial, and then each trial was baseline corrected and linear detrended. Finally, every trial was decimated to 100 Hz by applying a lowpass Chebyshev type I filter with a cutoff frequency of 40 Hz for antialiasing and then resampled by a factor of 10 in MATLAB R2012b (MathWorks, Natick, MA, USA).

Based on the preprocessed EEG data, 3D cortical distribution of current density with the properties of small localization bias and low spatial resolution was

determined using standardized LORETA (sLORETA) version 20081104 [11] as well as the Brainstorm 3.1 [12]. More precisely, a realistic head model was designed using the MNI152 template [13] as determined by the probabilistic Talairach atlas [14, 15] and symmetric boundary element method [16]. The standard electrode positions on the MNI152 scalp were adjusted with the fiducial points by manual inspection. The noise level in the EEG recordings prior to the aiming period was also estimated as the regularized full noise covariance matrix per subjects, so that the source reconstruction could be more accurate. Next, the intracerebral volume was partitioned in 15182 LORETA voxels at 4 mm spatial resolution so that sLORETA voxels could represent the electric activity at each dipole grids in neuroanatomic MNI space as the exact magnitude of the estimated current density with the signal to noise ratio of 3 dB. These 15182 voxels were then corrected to have an orientation that is close to the normal to the cortex, and finally 15028 voxels were estimated. In addition, the EEG cross-spectral matrix was computed to examine sLORETA voxels that generated the oscillatory activity in the delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz) rhythms.

Finally, the connectivity pattern analyses were performed to study the interconnectivity of the information processing elements between different cortical regions. The information processing units were calculated by averaging the source signals within a specific region of interest (ROI) for each orientation separately, and then taking the first component of the principal component analysis decomposition. The ROIs were defined based on 200 functionally distinct regions (100 ROIs in each hemisphere) using sources clustering method, in which each source was assigned to a single ROI in terms of closest distance to the center of mass of each ROI. The connectivity patterns between these ROIs were computed by means of two measures of N×N coherence and N×N phase locking value, and then such connectivity measures were averaged across subjects per each condition.

3 Results

EEG Localization. In the first step, a forward model of the head was computed to explain how cortical sources could influence the values on the EEG sensors. The estimated sources were standardized to minimize between-subject variability, and then averaged across subjects per each condition (Fig. 2). Considering PA as a reference case, relatively higher cortical source activations were more widely dispersed across cortical regions in C. Particularly, such activations were distributed apparently on both left fronto-temporal and right occipito-parieto-temporal areas during C, In addition, greater lateral temporal and lower medial central source activations respectively associated with alpha and beta rhythms were observed from the best who achieved minimum aiming variabilities in both conditions (1.9322 and 2.007 in PA and C) compared to the worst performer [4, 8], whose aiming variabilities were 4.0418 and 4.8819 in PA and C, respectively (Fig. 3).

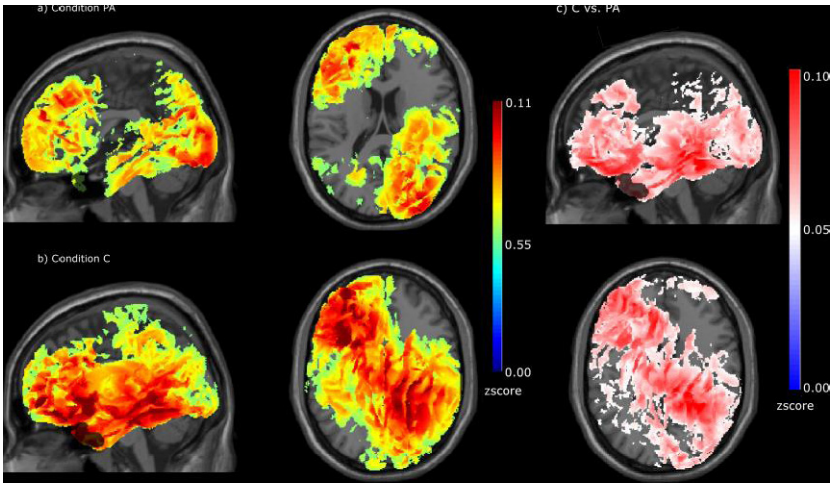


Fig. 2. Grand average EEG sources for both (a) PA and (b) C conditions respectively in each row. Each column displays the grand average of standardized cortical activations across subjects from left and top view, respectively; right view images are reversed as mirror images of left view images since each plot was drawn in 3D orthogonal sliced coordinates. (c) EEG source activations having $p < 0.10$ between two conditions were depicted.

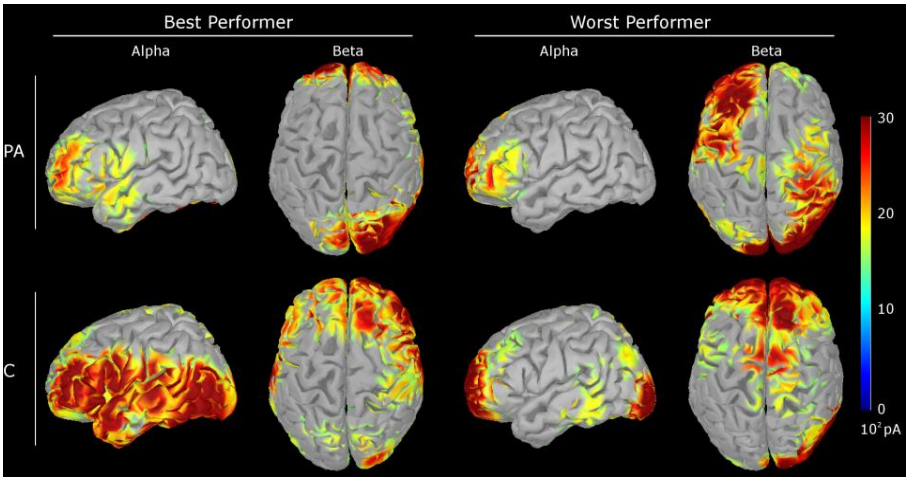


Fig. 3. Exemplar source activations in alpha (8-13 Hz; 1st and 3rd columns) and beta (13-30 Hz; 2nd and 4th columns) rhythms from best (1st and 2nd columns) and worst (3rd and 4th columns) performers under PA (1st row) and C (2nd row) conditions.

Brain Connectivity. Next, functional connectivity was computed from a clustered set of ROIs into prefrontal, frontal, temporal, parietal, and occipital regions in both left and right hemispheres (Fig. 4). Coherence results were similar in both conditions, but did reveal a tendency such that lower frequency bands were more globally connected

and higher frequency bands were more locally and anteriorly connected. Conversely, phase locking value could discriminate two conditions apparently in each frequency. Particularly, as similar to the sources distribution results, fronto-temporal and fronto-parietal connections associated with alpha rhythm were disconnected. Also, fronto-temporal, fronto-parietal, and occipito-parieto-temporal connections were attenuated in the beta frequency.

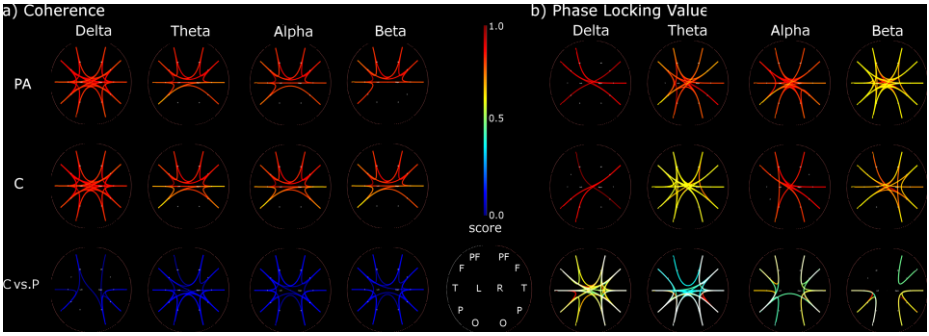


Fig. 4. Functional neural connectivity measures based on (a) N×N coherence and (b) N×N phase locking value for grand average ROIs per each condition. (c) All ROI groups were depicted in a polar grid, where each dot represented a distinct ROI. C-PA: C vs. PA, Left: Left hemisphere, Right: Right hemisphere, PF: Prefrontal, F: Frontal, T: Temporal, P: Parietal, O: Occipital regions.

4 Discussion

This study offered a multi-level examination of motor performance and cortical dynamics under competitive pressure. Previous studies have reported alpha power synchrony during expert marksmanship was positively related to performance and has been interpreted as quiescence of cognitive analysis during non-evaluative conditions [7, 10]. However, few studies have examined the impact of social evaluative mental stress on cortical dynamics during goal oriented motor behavior; to our knowledge, no studies exist where network connection analysis was performed to elucidate the functional relationship between brain regions under such competitive pressure condition. In the present study we examined how direct competition accompanied by a modest increase in mental stress perturbs cortical processes and influences the quality of motor performance.

Similar to our previous studies, the competitive condition increased the neural processing workload and resulted in heightened cortical activity across all of the topographical regions [10]. The elevation in cortico-cortical communication involved heightened connection between numerous non-motor regions with the motor planning region, suggesting a loss of psychomotor efficiency during social evaluation. The frontal input may be explained by elevated executive effort to inhibit task-irrelevant stimuli associated with the competition, while the central and parietal communication could be explained by additional effort in the motor and visuo-spatial domains.

Our results also indicated that competition produced behavioral changes in the fluency of motor performance, and source activations as well as functional connectivity were related to aiming variability, but no difference in shooting score (not shown in the results). Importantly examination of the best performer compared to the worst performer revealed a significant increase in left temporal alpha power during competition whereas the worst performer did not reveal a significant change in alpha. This is consistent with previous studies that have reported that the left temporal activity, associated with verbal-analytical processes, progressively decreases (reflected by increasing alpha synchrony in left temporal) during the aiming period of expert shooting up to the time of the trigger pull [17] and during the practice phase for motor skill acquisition [7]. Interestingly the best performer also did better during competition compared to performance alone, demonstrating the adaptive profile of alpha synchrony in left temporal region during competition [6] and supporting the notion of arousal-dependent performance facilitation to promote psychomotor efficiency [4]. In addition the lack of significant alpha synchrony in the left temporal region of the worst performer is consistent with the reinvestment hypothesis in which maladaptive self-talk (left temporal activation) infers with motor performance and results in neuromotor noise (elevated beta response in competition compared to performance alone).

Lastly the phase locking value results suggest competitive pressure can perturb the neural processes of the performer beyond that required simply to execute the pure motoric demands of a task because of the increase magnitude of the cortico-cotical communication across the alpha and beta frequency bands. Thus social evaluation may promote non-essential cortical activity, resulting in the degradation of motor efficiency in the form of nonessential limb movement (i.e., dysfluency of the aiming trajectory). Such a state could alter the motor preparatory processes (i.e., aiming) and the quality of the motor behavior, while the reduction in efficiency did not result in a change in performance outcome (as measure by shooting score).

In summary, the results revealed that competition introduced an increase in activity in the central nervous system, which introduced elevated non-essential neural activity to the visuo-motor processes, and then such a loss of psychomotor efficiency resulted in dysfluency of the aiming movement during competition. Since motor performance typically occurs under a variety of situations where workload demands and mental stress may perturb behavior, it appears useful to examine EEG functional connectivity and source distribution to aid in the assessment of performance optimization approaches and promote resilience to motor task inference.

References

1. Baumeister, R.F.: Choking Under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance. Journal of Personality and Social Psychology 46, 610–620 (1984)
2. Robazza, C., Pellizzari, M., Hanin, Y.: Emotion self-regulation and athletic performance: An application of the IZOF model. Psychology of Sport and Exercise 5, 379–404 (2004)

3. Deeny, S.P., Haufler, A.J., Saffer, M., Hatfield, B.D.: Electroencephalographic Coherence During Visuomotor Performance: A Comparison of Cortico–Cortical Communication in Experts and Novices. Journal of Motor Behavior 41, 106–116 (2009)
4. Rietschel, J.C., Goodman, R.N., King, B.R., Lo, L.-C., Contreras-Vidal, J.L., Hatfield, B.D.: Cerebral cortical dynamics and the quality of motor behavior during social evaluative challenge. Psychophysiology 48, 479–487 (2011)
5. Hatfield, B.D., Hillman, C.H.: The Psychophysiology of Sport: A Mechanistic Understanding of the Psychology of Superior Performance. In: Singer, R.N., Hausenblas, H.A., Janelle, C. (eds.) Handbook of Sport Psychology, pp. 362–386. John Wiley & Sons, New York (2001)
6. Hatfield, B.D., Haufler, A.J., Hung, T.-M., Spalding, T.W.: Electroencephalographic Studies of Skilled Psychomotor Performance. Journal of Clinical Neurophysiology 21, 144–156 (2004)
7. Kerick, S.E., Douglass, L.W., Hatfield, B.D.: Cerebral Cortical Adaptations Associated with Visuomotor Practice. Medicine and Science in Sports and Exercise 36, 118–129 (2004)
8. Hatfield, B.D., Costanzo, M.E., Goodman, R.N., Lo, L., Oh, H., Rietschel, J.C., Saffer, M., Bradberry, T., Contreras-Vidal, J.L., Haufler, A.J.: The Influence of Social Evaluation on Cerebral Cortical Activity and Motor Performance: A Study of "Real-Life" Competition (submitted for publication)
9. Masters, R., Maxwell, J.: The theory of reinvestment. International Review of Sport and Exercise Psychology 1, 160–183 (2008)
10. Rietschel, J.C., Miller, M.W., Gentili, R.J., Goodman, R.N., McDonald, C.G., Hatfield, B.D.: Cerebral-cortical networking and activation increase as a function of cognitive-motor task difficulty. Biological Psychology 90, 127–133 (2012)
11. Pascual-Marqui, R.D.: Standardized low resolution brain electromagnetic tomography (sLORETA): technical details. Methods and Findings in Experimental and Clinical Pharmacology 24(suppl.D), 5–12 (2002)
12. Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M.: Brainstorm: AUser-Friendly Application forMEG/EEG Analysis. Computational Intelligence and Neuroscience 2011, 879716 (2011)
13. Jurcak, V., Tsuzuki, D., Dan, I.: 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems. NeuroImage 34, 1600–1611 (2007)
14. Fuchs, M., Kastner, J., Wagner, M., Hawes, S., Ebersole, J.S.: A standardized boundary element method volume conductor model. Clinical Neurophysiology 113, 702–712 (2002)
15. Lancaster, J.L., Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., Kochunov, P.V., Nickerson, D., Mikiten, S.A., Fox, P.T.: Automated Talairach Atlas Labels For Functional Brain Mapping. Human Brain Mapping 10, 120–131 (2000)
16. Kybic, J., Clerc, M., Abboud, T., Faugeras, O., Keriven, R., Papadopoulo, T.: A Common Formalism for the Integral Formulations of the Forward EEG Problem. IEEE Transactions on Medical Imaging 24, 12–28 (2005)
17. Kerick, S.E., McDowell, K., Hung, T.-M., Santa Maria, D.L., Spalding, T.W., Hatfield, B.D.: The role of the left temporal region under the cognitive motor demands of shooting in skilled marksmen. Biological Psychology 58, 263–277 (2001)

Removal of Ocular Artifacts from EEG Using Learned Templates

Max Quinn[1], Santosh Mathan[2], and Misha Pavel[1]

[1] Oregon Health and Science University, Portland, OR 97239, USA
quinnma2013@alumni.ohsu.edu, pavelm@ohsu.edu
[2] Honeywell Laboratories, Redmond, WA 98052, USA
santosh.mathan@honeywell.com

Abstract. Electroencephalogram (EEG) data can provide information on cognitive states and processes with high temporal resolution, but to take full advantage of this temporal resolution, common transients such as blinks and eye movements must be accounted for without censoring data. This can require additional hardware, large amounts of data, or manual inspection. In this paper we introduce a greedy, template-based method for modeling and removing transient activity. The method iteratively models an input and updates a template; a process which quickly converges to a unique and efficient approximation of the input. When combined with standard source separation techniques such as Independent Component Analysis (ICA) or Principal Component Analysis (PCA), the method shows promise for the automatic and data driven removal of ocular artifacts from EEG data. In this paper we outline our method, provide evidence for its effectiveness using synthetic EEG data, and demonstrate its effect on real EEG data recorded as part of a minimally constrained cognitive task.

Keywords: EEG, EOG, ICA, PCA, BCI, matching pursuit.

1 Introduction

1.1 EOG Removal from EEG for Cognitive State Estimation

EEG is being actively investigated as a method for creating advanced human-computer interfaces, in which cortical activity patterns are used to infer the cognitive state of users or to control external devices.

Cognitive state estimation systems allow for estimates of latent cognitive properties such as mental workload, estimates which can then be used to regulate information provided to a user as part of an augmented cognition system, or for the development of new methods for assessing specific cognitive deficits and tailoring cognitive rehabilitation. Brain-computer interface (BCI) systems may allow patients with motor disabilities to interact with the world by controlling external devices through cortical activity. Devices include traditional computer control devices (such as mice), text communication systems, and robotic limbs.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 371–380, 2013.

In both cases, granular measures and high temporal resolution are desirable factors, as they allows for the production of more sensitive control mechanisms and a more detailed investigation of cognitive processing. Temporal resolution is easily hindered by physiological contamination from ocular and muscular activity. Recorded blinks and eye movements can be particularly disruptive due to their high amplitude and impulse-like shapes. These impulses disrupt both segment comparisons and local frequency estimates common in BCI and cognitive state estimation systems. The frequency with which these events occur makes the development of automatic systems for removing artifacts (rather than identifying and rejecting them) an important step in developing new tools built on EEG.

1.2 Existing EOG Removal Methods

A 2007 review by Fatourechi, et al. documents the common methods for removing such contaminants from EEG in BCI systems, their benefits and shortcomings, and the frequency with which they are employed in published work [1]. Although most studies neglect to discuss how ocular artifacts are treated (53.7%), the majority of those that employ automatic ocular artifact removal employ an electrooculogram (EOG) paired with a simple linear model (69.7%). This combination consists of dedicated electrodes placed around the eyes and a linear regression model that relates some portion of the signal recorded by each electrode of the EEG to an EOG component. In addition to requiring additional hardware and a more involved setup process, the relationship between the recorded EOG signal and the EEG is not entirely linear. The activity associated to the EOG during different eye movements varies depending on the type of eye-movement [2].

The second most common method for ocular component removal, being reported in 9.1% of papers, is blind source separation, usually in the form of independent component analysis (ICA). ICA is a statistical method that separates additively mixed, independent signals through the optimization of a measure of non-normality in the resulting component signals. ICA may be used with or without an EOG [3,4], and depends on both a large amount of data (for the optimization process) and a manual identification step (to identify components associated with the type of contamination in question).

Only 6.1% of automatic EOG artifact removal is performed with principal component analysis (PCA), an eigendecomposition based method for identifying uncorrelated components. The rarity with which PCA is used is probably due to EOG components and cortical activity not meeting the condition that components be orthogonal. PCA decompositions will produce components that contain combinations of activity from unrelated sources (such as ocular activity and frontal cortical activity). Yet, PCA remains a tempting method because the decomposition is deterministic, fast, and requires less data (when compared to ICA).

ICA and PCA perform a similar task, using statistical properties to identify spatial components that have some sort of coherent activation pattern. However, they differ in terms of assumptions about the data and the difficulty with which projections can be derived from the data. PCA alone is a poor match for the removal of EOG artifacts due to an orthogonality assumption not met by EEG data [5]. However, dropping this assumption and further optimizing for independence makes ICA a more difficult method to employ, in terms of data requirements and derivation properties. For both methods, it is likely the case that the estimated EOG component also contains cortical activity, possibly due to unmet requirements for full separation under PCA, and to suboptimal projections being derived through the optimization for non-normality [5,6].

1.3 Using Temporal Shape for Artifact Removal

In this paper, we describe an extension of ICA-based or PCA-based EOG artifact removal that further constrains the attribution of signal to ocular activity by leveraging the conserved temporal shape of common transients. Our method uses a modified version of matching pursuit, a method by which signals are represented as a linear combination of elements from an over-complete dictionary [8]. Matching pursuit results in a sparse signal representation which has been found to be useful in signal classification and compression contexts. However, the modeling process is prohibitively slow when using common dictionaries, making it unsuitable for online applications [9]. Although applied to EEG soon after its development [7], matching pursuit is rarely applied in the context (not appearing in Fatourechi's survey of artifact removal techniques) due to this poor runtime performance. Methods for accelerating matching pursuit rely on efficient implementation and careful pruning of the dictionary elements considered during signal modeling. Although the gains from these approaches can be substantial, the runtime remains bound to the cardinality of the dictionary being used, which is usually large [9]. We find that artifacts in EEG are sufficiently modeled using a very simple dictionary, consisting of only a few entries. Our approach modifies the matching pursuit process by comparing dictionary elements to the signal at all possible offsets, reducing the dictionary size to only a few elements that can be learned from the data, and are constructed to model an individual subjects ocular artifacts for removal.

Evaluation of EEG processing methods presents an inherent challenge, as we lack a ground truth signal against which processed signals can be compared. For this reason, processing methods are sometimes evaluated using synthetic data, where conclusions regarding a method can only be drawn in so far as the synthetic data is a meaningful approximation of real data. We will demonstrate our method in this synthetic context first, where assumptions will be made explicit and efficacy can be demonstrated clearly, and then we will demonstrate the method using real EEG data, which can not be evaluated for correctness, but compares favorably to the synthetic results.

2 Methods

2.1 Data Collection

We developed our technique for removing ocular artifacts while investigating EEG data collected during performance of a naturalistic reading task. Subjects read passages under a high-workload condition, where text came from sources such as the New Yorker and a challenging time constraint was imposed, or under a low-workload condition, which used easier passages and little time constraint. The nature of the task made it unreasonable to discourage eye-movements, and frequency with which they occurred made rejection unreasonable, so we had to remove the EOG from the recorded data.[1]

The frequency of eye movements made ICA quite effective in producing an independent ocular component using only a few minutes of recorded data. The component was found and identified using the methods described in Jung, et al. [4]. To further isolate ocular activity, we modeled the activation time-course of this channel using our template based method. The modeled ocular activity was then subtracted from this component, leaving other activity as a residual that could be re-integrated into the data at large. A mathematical description of the modeling process follows.

2.2 Processing

The component recognized as containing ocular activity, $\mathbf{x} = (x_1, x_2, ..., x_n)$, is modeled using a shape template $\mathbf{h} = (h_1, h_2, ..., h_m)$ and a scaling coefficient sequence $\mathbf{c} = (c_1, c_2, ..., c_n)$.

The signal model $\mathbf{m}$ is constructed by the convolution $\mathbf{m} = \mathbf{h} * \mathbf{c}$. From the signal $\mathbf{x}$ and this model $\mathbf{m}$, we generate the residual $\mathbf{r} = \mathbf{x} - \mathbf{m}$.

For convenience, let $N_m(x_i) = (x_{i-\frac{m}{2}}, ..., x_{i+\frac{m}{2}})$, that is, an m unit neighborhood of x_i. We initialize $\mathbf{h}$ by taking the point-wise average of signal segments that are centered on local amplitude extrema.

$$W = \{w_i : max(|N_m(x_i)|) = |x_i|\},$$
$$\text{where } w_i = \frac{N_m(x_i)}{stdev(N_m(x_i))}$$
$$\mathbf{h} = \mathrm{E}(W)$$

The coefficient sequence $\mathbf{c}$ is initialized with $c_i = 0$ for all i. We greedily add single coefficients using the following update procedure.

$$i^* = \arg\max_i cov(\mathbf{h}, N_m(r_i))$$

[1] EEG data was recorded from standard scalp locations using a 64-channel Biosemi™ ActiveTwo™ system. Data was down-sampled to 256Hz during recording and was high-pass filtered at 1 Hz using an 8th order Butterworth filter. Additional details regarding the experiment and the collection of this data can be found in the Engineering in Medicine and Biology 2010 conference proceedings [10].

$$a^* = \arg\min_a \sum (a\mathbf{h} - N_m(r_{i^*}))^2$$

$$c_{i^*} = a^*$$

$$\text{Update } \mathbf{m} = \mathbf{h} * \mathbf{c} \text{ and } \mathbf{r} = \mathbf{x} - \mathbf{m}$$

Each iteration selects a location from the current residual, fits a scaling coefficient, and updates the model and residual. These steps are repeated until a termination condition is met, which can be based on a limit on the number of non-zero elements of $\mathbf{c}$, a threshold on a^*, or reaching a target residual variance.

After the signal has been modeled, we update the template.

$$\text{Let } \mathbf{r}_i^\circ = \mathbf{x} - \mathbf{h} * (c_1, c_2, ..., c_{i-1}, 0, c_{i+1}, ..., c_n),$$

$$W = \{N_m(r_i^\circ) : c_i \neq 0\},$$

$$\mathbf{h} = \mathrm{E}(W).$$

$\mathbf{r}_i^\circ$ shows us our current residual modified such that a single transient remains unmodeled. This allows us approximate the event at i in isolation, reducing the effect of overlapping occurrences of the transient.

Repeating this process of modeling and template updating based on covariance guides the model toward a group of transients with a similar shape, producing a template increasingly fitting the transient of interest. A final signal estimate $\mathbf{m}_{final}$ is built using the same matching pursuit process that is used during each of the update steps. Additional transients may be modeled by repeating the entire process, starting with $\mathbf{x} - \mathbf{m}_{final}$.

3 Results

3.1 Synthetic Data

To evaluate how well our method for modeling transients decomposes an EEG signal containing both cortical and EOG components, we generated synthetic EEG data with an additive combination of a stable random signal and a sequence of transients generated from a conserved temporal shape. That is, $\mathbf{x_{EEG}} = \mathbf{x_{stable}} + \sum_i \mathbf{h_i} * \mathbf{c_i}$, where $\mathbf{h_i}$ is a transient, and $\mathbf{c_i}$ is a sparse activation sequence for $\mathbf{h_i}$.

Transients are often impulse like (and therefor broadband), so their frequency distributions overlap with that of the stable signal. We generated a transient by producing a random .5 second signal in the 1-8Hz band, and scaling it to have unit variance. There were 30 activations with coefficients in the range .75 to 1.25. The random signal was from the 1Hz to 20Hz frequency range and had variance of .25. This process produced the signal seen in Figure 1. Although not a sophisticated model of EEG activity, the components overlap in frequency distribution, and are reminiscent of contaminated EEG components.

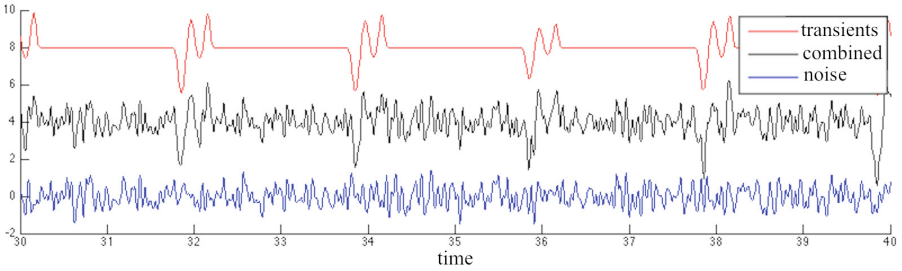


Fig. 1. We model EEG data as the additive combination of a stable random signal between 1 and 20 Hz and repeated activations of a transient in the range 1 to 8 Hz. Our model includes substantial activity not associated with the transient, which would indicate poor separation of signal components when applying a source separation technique to real EEG data.

Initializing the template using signal segments of high local variance produces a fairly poor result. However, applying our template updating procedure several times produces a reasonable approximation, as can be seen in Figure 2. Using the resulting template approximation, the signal is separated into a transient activation sequence and a residual signal, shown in Figure 3

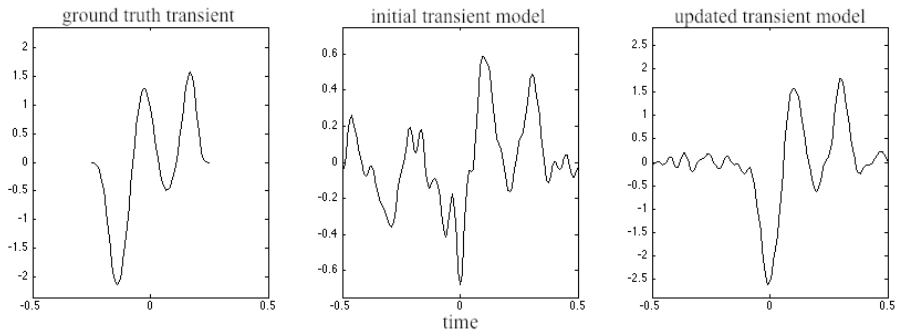


Fig. 2. Our method for modeling a repeated transient produces a good approximation of the generating transient signal, despite a poor initial estimate. The ground truth transient shape is shown on the left, the initial estimate is shown in the middle, and the final approximation is shown on the right.

As we iterated between signal modeling and template updating, we see in Figure 4 that the residual error, the squared sum of the difference between our model residual and the stable random signal, quickly drops off to a constant near the original variance. Although the error cannot be monitored during application to a real signal, the residual variance can be monitored. Figure 4 also shows that when the remaining residual variance stabilizes, there is also little change in the

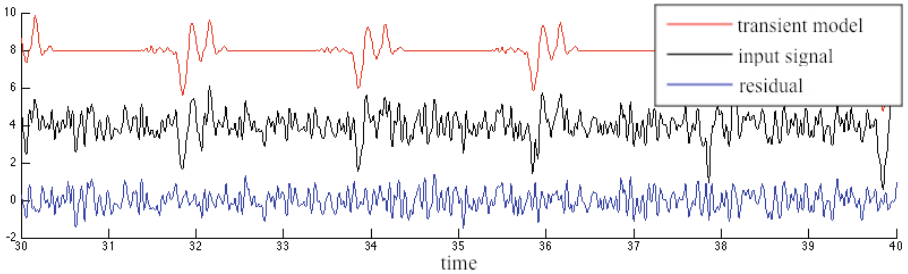


Fig. 3. The modeled signal produces a close approximation of the two contributing signal sources (seen in Figure 1) from a single channel. In conjunction with source separation techniques that use spatial information, our method may produce a more complete separation of contributing sources.

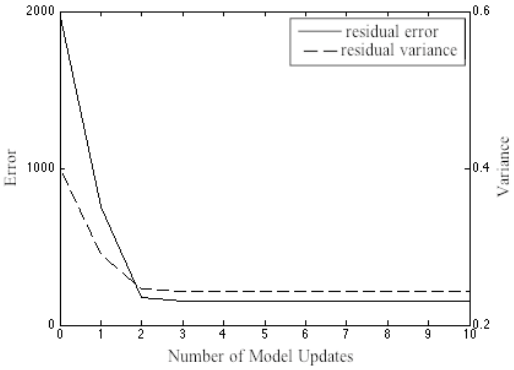


Fig. 4. The residual error closely tracks the total remaining variance in the signal. This provides useful information for constructing an appropriate termination condition when using real data, where error cannot be monitored.

residual error. If this relationship between residual variance and error remains consistent when using real data, it contributes to a natural termination condition for the signal modeling procedure.

3.2 EEG Data

To demonstrate our method for modeling ocular transients, we use a component from a PCA decomposition of the aforementioned data recorded during a reading task. Although our method works well in conjunction with ICA, PCA quickly produces a robust decomposition from which components associated with ocular activity might be more easily identified through automated processes. However, the ocular activity is less isolated when using PCA. Our template approach addresses this issue by introducing the additional shape constraint.

PCA was applied to our recorded EEG data. We selected the component containing blink activity based by examining the scalp distribution of PCA weights. Our signal modeling technique was applied to the the activation time-course of this component. As before, the template modeling method converges on a reasonable approximation of the time-course of the most common transient, as can be seen in Figure 5. Using this template, the activation time-course of the selected PCA component is decomposed into two parts: the modeled transient activation and the residual activity, which is presumably cortical in nature. This further decomposition can be seen in Figure 6. Again, the modeled signal closely tracks what appear to be instances of the transient activity, without introducing obvious artifacts or removing additional activity.

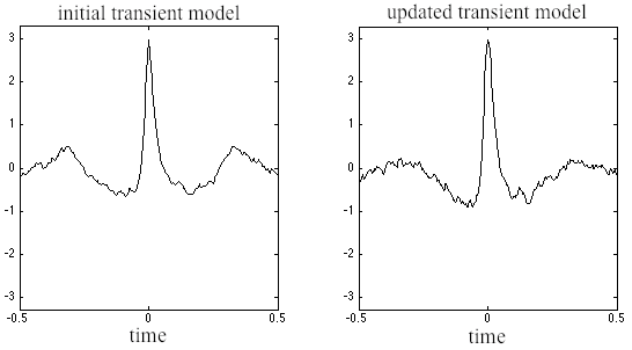


Fig. 5. The estimated transient converges quickly to a plausible shape for common ocular artifacts. The frequency of blinks in rapid succession contributes to an elevated amplitude leading and following the main amplitude spike by about .25 seconds. These sorts of artifacts associated with overlapping transients are reduced by the local modeling used when learning the transient shape.

Using this method allowed us to perform our analyses of data recorded during the reading-task, which included the extraction of local frequency features, without corruption from impulse-like activations.

4 Discussion

In this paper we introduced a template based approach to modeling transients in EEG data. We provided evidence for the efficacy of the method in the context of synthetic data, and demonstrated the result of applying the process to real EEG data. The method provides the benefit of being able to remove common ocular contaminants from recorded EEG signals without rejecting data; without the use of additional hardware; and, when combined with PCA, without requiring large amounts of data or a detailed analysis of separated components.

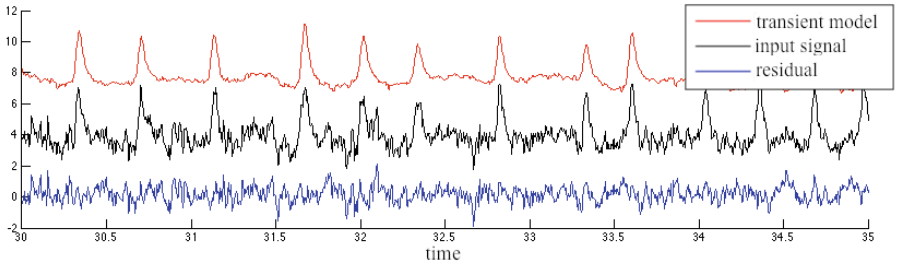


Fig. 6. On ocular and frontal activity component (in black) is split into a transient component (in red) and a locally stationary signal (in blue). It appears that this separation, using a component derived using PCA, creates for a more easily automated method for removing ocular contamination in EEG when compared to ICA based methods, which usually require the attention of a researcher to identify ocular components.

There are several natural concerns associated with the method, including implementation details such as termination conditions and the potential for learning degenerate templates that model more than the intended transient activity. Additionally, the known shortcomings of matching pursuit remain applicable, with the potential for highly suboptimal signal representations to occur when transients overlap frequently. Furthermore, we have only peripherally discussed the situation where multiple transient patterns occur within the same signal. For example, transients associated with left-to-right and right-to-left eye movements usually share a component when identified using PCA or ICA, but may not be appropriately modeled with a single template.

In practice, many of these concerns can be mitigated, without significant effort, by empirically adjusting the few parameters in the system, such as the width of the transient model and the threshold on activation coefficients. The presence of multiple transient shapes requires a bit more engineering, but initial results using clustering of signal segments (rather than a simple mean), or sequential application of the method, are encouraging.

Several additional experiments remain future work, including: comparing our method to EOG signals gathered with dedicated electrodes, further automation through the integration of priors for the spatial distribution of common ocular contaminants, and the modeling and detection of evoked response potentials without the use of a time-lock.

References

1. Fatourechi, M., Bashashati, A., Ward, R.K., Birch, G.E.: EMG and EOG Artifacts in Brain Computer Interface Systems: A Survey. Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology 118(3), 480–494 (2007)
2. Croft, R.J.: The Removal of Ocular Artifact from the EEG. University of Wollongong (1999)

3. Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P.: Recipes for the Linear Analysis of EEG. NeuroImage 28(2), 326–341 (2005)
4. Jung, T., Humphries, C., Makeig, S., Mckeown, M.J., Iragui, V., Sejnowski, T.J.: Extended ICA Removes Artifacts from Electroencephalographic Recordings. Neural Information Processing Systems 10, 894–900 (1998)
5. Jung, T., Humphries, C., Lee, T., Makeig, S., Mckeown, M.J., Iragui, V., Sejnowskil, T.J.: Removing Electroencephalographic Artifacts: Comparison Between ICA and PCA. Neural Networks for Signal Processing VIII. In: Proceedings of the 1998 IEEE Signal Processing Society Workshop, pp. 63–72 (1998)
6. Castellanos, N.P., Makarov, V.A.: Recovering EEG Brain Signals: Artifact Suppression with Wavelet Enhanced Independent Component Analysis. Journal of Neuroscience Methods 158, 300–312 (2006)
7. Durka, P.J., Blinowska, K.J.: Analysis of EEG Transients by means of Matching Pursuit. Annals of Biomedical Engineering 23(5), 608–611 (1995)
8. Mallat, S.G., Zhang, Z.: Matching Pursuits With Time-Frequency Dictionaries. IEEE Transactions on Signal Processing 41(12), 3397–3415 (1993)
9. Krstulovic, S., Gribonval, R.: Mptk: Matching Pursuit Made Tractable. In: ICASSP 2006 Proceedings, 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3 (2006)
10. Mathan, S., Smart, A., Ververs, T., Feuerstein, M.: Towards an Index of Cognitive Efficacy. Engineering in Medicine and Biology (2010)

Brain in the Loop Learning Using Functional Near Infrared Spectroscopy

Patricia A. Shewokis[1,2,3], Hasan Ayaz[1,2], Adrian Curtin[1,2], Kurtulus Izzetoglu[1,2], and Banu Onaral[1,2]

[1] School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, PA 19104, USA
[2] Cognitive Neuroengineering and Quantitative Experimental Research (CONQUER) Collaborative, Drexel University, Philadelphia, PA 19104, USA
[3] Nutrition Sciences Department, College of Nursing and Health Professions, Drexel University, Philadelphia, PA 19102, USA
{shewokis,hasan.ayaz,abc48,ki25,bo26}@drexel.edu

Abstract. The role of practice is crucial in the skill acquisition process and for assessments of learning. In this study, we used a portable neuroimaging technique, functional near infrared (fNIR) spectroscopy for monitoring prefrontal cortex activation during learning of spatial navigation tasks throughout 11 days of training and testing. Two different tasks orders, blocked and random, were used to test the effect of the practice schedule on the acquisition and transfer of 3D computer mazes. Results indicated variable decreases in the hemodynamic response during the initial days of practice. Although there were no differences in mean oxygenation for the practice orders across acquisition the random practice order used less oxygenation than the blocked order for the more difficult tasks in the transfer phase Use of brain activation and behavioral measures provides can provide a more accurate depiction of the learning process. Since fNIR systems are safe, portable and record brain activation in ecologically valid settings, fNIR can contribute to future learning settings for assessment and personalization of the training regimen.

Keywords: Optical Brain Imaging, functional near infrared spectroscopy, fNIR, Learning, Spatial navigation, contextual interference.

1 Introduction

The advent of new and improved brain imaging tools, that allow monitoring brain activity in ecologically valid environments, is expected to allow better identification of neurophysiological markers of human performance and learning. Further, deployment of portable neuroimaging technologies to real time settings could help assess cognitive and motor task related brain activations for objective assessment of mental effort and cortical processing involved for the task at hand. Functional Near-Infrared Spectroscopy (fNIR) is an emerging optical brain imaging technology that relies on

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 381–389, 2013.

optical techniques to detect changes of hemodynamic responses within the prefrontal cortex in response to sensory, motor, or cognitive activation.

The role of practice is crucial in the skill acquisition process and for assessments of learning. By examining the cognitive and behavioral output during the performance and learning of selected cognitive and motor tasks, along with a detailed examination of the neural activity obtained from fNIR, it may be possible to gain insight into the impact of practice on learning, transfer and the skill acquisition processes. This paper discusses the neural mechanisms of learning and skill acquisition using fNIR with Maze Suite 3D spatial navigation tasks using a contextual interference paradigm.

The organization of practice when learning multiple tasks (e.g., [1, 2]) is a learning phenomenon called the contextual interference effect. The effects of contextual interference are evident when individuals acquire multiple tasks under different practice schedules. High contextual interference (random (RAN) practice order) is created when the tasks to be learned are presented in a non-sequential, unpredictable order. Low contextual interference (blocked (BLK) practice order) is created when the tasks to be learned are presented in a predictable order

The specific aim of this pilot study is to identify brain based biomarkers of learning and its relationship to task performance improvement with practice as measured by fNIR spectroscopy which is a safe, non-invasive, affordable and portable neuroimaging technology that can be used to monitor hemodynamic changes that occur in the brain, i.e., blood oxygenation and blood volume, during select cognitive tasks such as mental workload [3-6], task difficulty/problem solving [7-9], performance[10-12] and learning[12-14] assessment tasks. Moreover, fNIR data can be collected in quiet settings unlike functional magnetic resonance imaging (fMRI) that exposes subjects to noise and confines them to restricted spaces and a supine position during the data acquisition process. These qualities pose fNIR as an ideal methodology for monitoring cognitive activity-related hemodynamic changes not only in laboratory settings but also under ecologically valid conditions – real world environments.

For the experimental paradigm, fNIR measures were integrated into a virtual 3D navigation tasks generated with MazeSuite [15, 16] (Drexel University). The protocol involved execution of wayfinding tasks throughout 11 days. Two different groups, BLK and RAN practice orders were used for learning of mazes (virtual environments /labyrinths) during acquisition and more difficult (complex) mazes during retention. A 16-channel continuous wave (CW) fNIR system designed by the Optical Imaging Team at Drexel University (see [3]) was used to monitor the prefrontal cortex during task performance.

2 Methods

2.1 Participants

Eight right-handed participants (assigned using the Edinburgh Handedness Inventory[17]) volunteered for this study. Participants self-reported that they did not have any neurological or psychiatric history; that they were medication-free, and had

normal or corrected-to-normal vision. Participants gave written informed consent for the study, which was approved by the Institutional Review Board at Drexel University, and were paid for their participation. Participants were randomly assigned to either a BLK practice order or RAN order.

2.2 Experiment Protocol

The spatial navigation tasks involved wayfinding in virtual 3D environments rendered using MazeSuite software [15, 16] developed in our lab. Figure 1 below displays a screen from a one of the 3D maze (labyrinths) that participants interacted with using keyboard and mouse controls. The first day of the experiment involved familiarization with the task controls and generic navigation in an orientation maze. Tasks for the acquisition period (3 mazes) were performed on each day 2 through day10. On day 11, transfer tasks (2 novel mazes) were executed. For the BLK group, one type of maze was practiced on each day with three days of practice per maze. For the RAN group, all mazes were practices on all days. Total of mazes for all subjects were same (acquisition: 9 days x 15 repetitions per day + transfer: 1 day x 12 repetitions per day). Transfer practice order was the same as the acquisition order with the BLK group having the transfer mazes in a blocked order while the RAN group had the transfer mazes in a random order. The transfer mazes were used to determine the extent to which each subject was able to generalize their learning and practice with acquisition mazes – given that robust learning assessments are best illustrated through generalizability tests like transfer.

Fig. 1. Functional Near Infrared Spectroscopy sensor (head band) covers forehead of participants (left) and screen shot from a maze rendering on computer screen (right).

2.3 fNIR Data Acquisition

The continuous wave fNIR system (fNIR Devices LLC; www.fnirdevices.com) used in this study is connected to a flexible sensor pad that contains 4 light sources with built in peak wavelengths at 730 nm and 850 nm and 10 detectors designed to sample cortical areas underlying the forehead. With a fixed source-detector separation of 2.5 cm, this configuration generates a total of 16 measurement locations (optodes) [3, 18]. For data acquisition and visualization, COBI Studio software [15] (Drexel University) was used. The sampling rate of the system was 2Hz. During the task, a serial cable

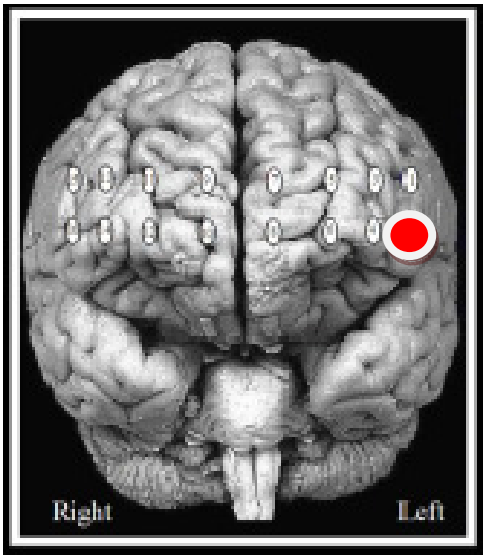


Fig. 2. Measurement locations of the 16 optodes [18]. The location of optode #2 (indicated by the red circle) is close to AF7 in the International 10–20 System and is located within the left prefrontal cortex (inferior frontal gyrus).

between the fNIR data acquisition computer and MazeSuite presentation computer was used to transfer time synchronization signals (markers) that indicate the start of sessions and onset of maze tasks.

2.4 Data Analysis

For each participant, raw fNIR data was low-pass filtered with a finite impulse response, linear phase filter with order of 20 and cut-off frequency of 0.1Hz to attenuate the high frequency noise [3]. Motion artifact contaminated sessions and saturated channels (if any), in which light intensity at the detector was higher than the analog-to-digital converter limit were excluded [19]. Using time synchronization markers, fNIR data segments for rest periods (15 seconds rest period between trials) and task periods (maze task performance) were extracted. Blood oxygenation changes within dorsolateral prefrontal cortex for all optodes were calculated using the Modified Beer Lambert Law (MBLL) for task periods with respect to rest periods at beginning of each task[3]. Dependent measures included relative changes in the mean oxygenation change for optode #2 (see Fig 2) and behavioral measure of path length for the mazes For acquisition for optode #2 mean oxygenation and mean path length, 2 X 9 (Practice Order X Day) mixed model ANOVAs with repeated measures on the last factor. In this repeated measures design, participants were considered a random-effects factor, whereas Practice Order was considered a fixed-effect factor. To test a fixed-effect with one random effect in the model, the appropriate denominator term for the

F-statistic was determined by limiting the error term for the interaction of the fixed and random factors to zero [20]. For transfer, planned contrasts of the BLK vs RAN practice orders were calculated for optode #2 mean oxygenation changes and mean path length. The significance criterion for all tests was set at $\alpha = 0.05$.

2.5 Behavioral Measures

For acquisition, the behavioral measure mean path length (arbitrary units (a.u.), had a significant interaction of Practice Order by Day with [$F_{(8,920)} = 7.43$, $p < 0.001$] and significant main effect of Day [$F_{(8,920)} = 22.82$, $p < 0.001$]. The main effect of Practice Order was not significant [$F_{(1,920)} = 3.14$, $p = 0.137$]. The change of average path length for the BLK and RAN groups across acquisition and transfer is depicted in Fig. 3. The planned contrast resulted in no significant difference between the BLK and RAN practice orders in the transfer phase with [$F_{(1,94)} = < 1.0$, $p = 0.591$].

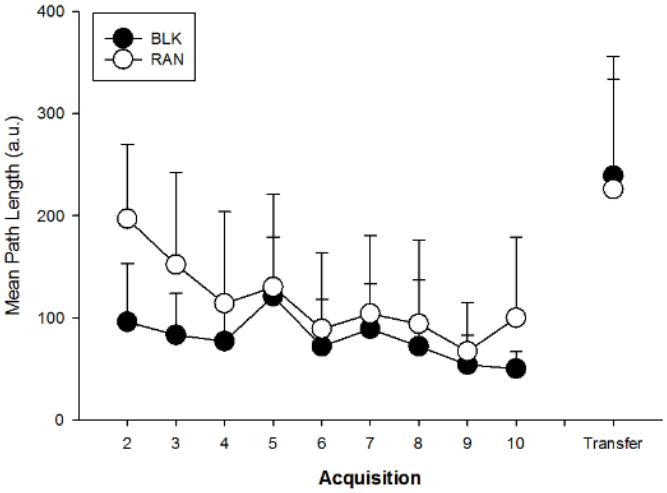


Fig. 3. Average navigation path length for all subjects during acquisition and transfer for blocked order group (BLK) and random order group (RAN). Error bars are standard deviations (SD)

2.6 fNIR Measures

In this paper, the left inferior frontal gyrus (location of optode #2 – see Fig. 2) mean oxygenation change (μmolar) values were assessed for all maze trials across the acquisition phase (9days) and for the transfer phase (day 11) transfer. Both the interaction of Practice Order X Days [$F_{(8,991)} = 2.03$, $p=0.04$], and the main effect of Days [$F_{(8,991)} = 2.00$, $p =0.043$] were significant for acquisition. There was no significant main effect of Practice Order with [$F_{(1,991)} = < 1.0$, $p =0807$]. For transfer, the planned contrasts yielded a significant difference with [$F_{(1,84)} = 6.86$, $p =0.01$]. Depicted in

Fig.4 are the mean oxygenation changes for practice orders plotted as a function of the acquisition and transfer phases. More difficult tasks were performed during transfer and the change in oxygenation values was higher for the BLK practice order relative to the end of acquisition. However, oxygenation for the RAN practice order was lower compared to the BLK practice order during transfer and the RAN practice order had lower mean oxygenation during transfer relative to the end of acquisition.

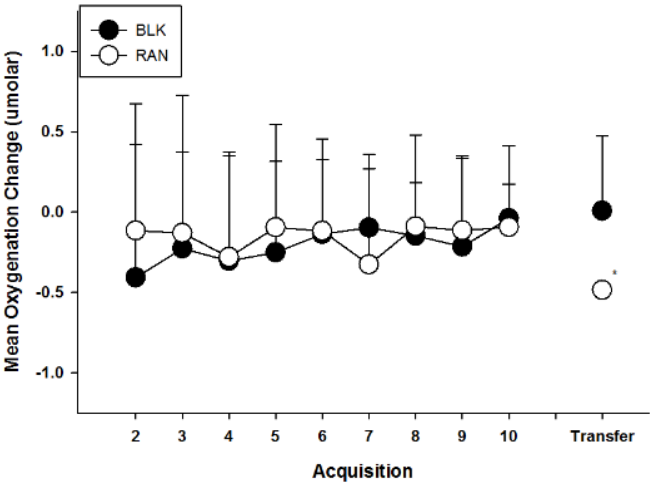


Fig. 4. Average oxygenation changes at optode #2 across acquisition and transfer stages. Error bars are standard deviations (SD). * ($p < 0.05$).

3 Discussion

The purpose of this study was to test the practice order effect with spatial navigation tasks. Behavioral performance measures and cortical hemodynamic responses as measured by wearable optical brain imaging were collected to compare changes across 11 days of practice and for two practice orders: BLK and RAN. Our results indicate differential patterns for behavioral and fNIR measures for the different practice orders.

During acquisition phase (day2-10), navigation path length (behavioral measure) improved across both practice orders for the acquisition phase (Fig. 3). This expected results showed that participants improved their navigation skill with more practice. The mean path length traveled was lower for BLK order across the days of the acquisition period compared to the RAN practice order. The oxygenation (fNIR measure) was variable throughout the acquisition phase for both practice orders (Fig. 4). There was a quicker reduction in oxygenation for the BLK practice order relative to the RAN order and both BLK and RAN orders had similar oxygenation values at the end of acquisition.

During the transfer phase (day11), more complex maze tasks were presented. Average navigation path length for RAN group was higher compared to BLK group, suggesting that RAN practice order prepared the participants for the more complex task. Similarly, oxygenation during the transfer phase was lower for the RAN group compared the BLK group suggesting that RAN group used less mental effort to complete the task compared to the BLK order group. These findings corroborate the PET findings with spatial navigation of virtual mazes reported by Van Horn and colleagues [21]. In addition, using fMRI, Wymbs and Grafton [22] reported that the left inferior frontal gyrus was differentially activated during late learning as a function of practice schedule for the sequence execution of a go/no-go task. Our transfer findings illustrate that there is a differential relative mean oxygenation of the left inferior frontal gyrus region for RAN and BLK practice orders for spatial navigation tasks. These results help to extend our understanding of the contextual interference effect regarding the influences of the practice order and task type on neural function [21-26].

This study tested the effects of learning spatial navigation tasks in virtual environments. Results indicated that behavioral performance and oxygenation in the anterior prefrontal cortex is sensitive to both the amount of practice and the order of practice in learning multiple tasks. This study provides preliminary information about fNIR measures of the anterior prefrontal cortex hemodynamic response and its relationship to learning/skill acquisition. Since fNIR technology allows the development of mobile, non-intrusive and miniaturized devices, it has the potential to be used in future learning/training environments to provide objective, task related brain-based measures for optimizing the learning process.

Acknowledgement. This study was funded in part under a U.S. Army Medical Research Acquisition Activity; Cooperative Agreement W81XWH-092-0104. The content of the information herein does not necessarily reflect the position or the policy of the U.S. Government or the U.S. Army and no official endorsement should be inferred.

References

1. Magill, R.A., Hall, K.G.: A review of the contextual interference effect in motor skill acquisition. Human Movement Science 9, 241–289 (1990)
2. Shewokis, P.A.: Memory consolidation and contextual interference effects with computer games. Perceptual and Motor Skills 97, 581–589 (2003)
3. Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B.: Optical brain monitoring for operator training and mental workload assessment. Neuroimage 59, 36–47 (2012)
4. Girouard, A., Solovey, E.T., Jacob, R.J.K.: Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. International Journal of Autonomous and Adaptive Communications Systems 6, 26–44 (2013)
5. James, D.R.C., Orihuela-Espina, F., Leff, D.R., Sodergren, M.H., Athanasiou, T., Darzi, A.W., Yang, G.Z.: The ergonomics of natural orifice translumenal endoscopic surgery (NOTES) navigation in terms of performance, stress, and cognitive behavior. Surgery 149, 525–533 (2011)

6. James, D.R.C., Orihuela-Espina, F., Leff, D.R., Mylonas, G.P., Kwok, K.-W., Darzi, A.W., Yang, G.-Z.: Cognitive burden estimation for visuomotor learning with fNIRS. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part III. LNCS, vol. 6363, pp. 319–326. Springer, Heidelberg (2010)
7. Ayaz, H., Shewokis, P.A., İzzetoğlu, M., Çakır, M.P., Onaral, B.: Tangram solved? Prefrontal cortex activation analysis during geometric problem solving. In: 34th Annual International IEEE EMBS Conference, pp. 4724–4727. IEEE (2012)
8. Çiftçi, K., Sankur, B., Kahya, Y.P., Akın, A.: Functional Clusters in the Prefrontal Cortex during Mental Arithmetic. In: 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, pp. 1–4 (2008)
9. Hampshire, A., Thompson, R., Duncan, J., Owen, A.M.: Lateral prefrontal cortex subregions make dissociable contributions during fluid reasoning. Cerebral Cortex 21, 1–10 (2011)
10. Ayaz, H., Bunce, S., Shewokis, P., Izzetoglu, K., Willems, B., Onaral, B.: Using Brain Activity to Predict Task Performance and Operator Efficiency. In: Zhang, H., Hussain, A., Liu, D., Wang, Z. (eds.) BICS 2012. LNCS, vol. 7366, pp. 147–155. Springer, Heidelberg (2012)
11. Power, S.D., Kushki, A., Chau, T.: Towards a system-paced near-infrared spectroscopy brain-computer interface: differentiating prefrontal activity due to mental arithmetic and mental singing from the no-control state. Journal of neural engineering 8, 066004 (2011)
12. Ayaz, H., Cakir, M.P., Izzetoglu, K., Curtin, A., Shewokis, P.A., Bunce, S.C., Onaral, B.: Monitoring expertise development during simulated UAV piloting tasks using optical brain imaging. In: 2012 IEEE Aerospace Conference, pp. 1–11 (2012)
13. Shewokis, P.A., Ayaz, H., Izzetoglu, M., Bunce, S., Gentili, R.J., Sela, I., Izzetoglu, K., Onaral, B.: Brain in the Loop: Assessing Learning Using fNIR in Cognitive and Motor Tasks. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2011. LNCS, vol. 6780, pp. 240–249. Springer, Heidelberg (2011)
14. Pfurtscheller, G., Bauernfeind, G., Wriessnegger, S.C., Neuper, C.: Focal frontal (de)oxyhemoglobin responses during simple arithmetic. Int. J. Psychophysiol. 76, 186–192 (2010)
15. Ayaz, H., Shewokis, P.A., Curtin, A., Izzetoglu, M., Izzetoglu, K., Onaral, B.: Using MazeSuite and Functional Near Infrared Spectroscopy to Study Learning in Spatial Navigation. J. Vis. Exp., e3443 (2011)
16. Ayaz, H., Allen, S.L., Platek, S.M., Onaral, B.: Maze Suite 1.0: a complete set of tools to prepare, present, and analyze navigational and spatial cognitive neuroscience experiments. Behav Res. Methods 40, 353–359 (2008)
17. Oldfield, R.C.: The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113 (1971)
18. Ayaz, H., Izzetoglu, M., Platek, S.M., Bunce, S., Izzetoglu, K., Pourrezaei, K., Onaral, B.: Registering fNIR data to brain surface image using MRI templates. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 2671–2674 (2006)
19. Ayaz, H., Izzetoglu, M., Shewokis, P.A., Onaral, B.: Sliding-window Motion Artifact Rejection for Functional Near-Infrared Spectroscopy. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., pp. 6567–6570 (2010)
20. Maxwell, S., Delaney, H.: Designing experiments and analyzing data: A model comparison perspective. Psychology Press, NY (2004)
21. Van Horn, J.D., Gold, J.M., Esposito, G., Ostrem, J.L., Mattay, V., Weinberger, D.R., Berman, K.F.: Changing patterns of brain activation during maze learning. Brain Res. 793, 29–38 (1998)

22. Wymbs, N.F., Grafton, S.T.: Neural substrates of practice structure that support future off-line learning. Journal of Neurophysiology 102, 2462–2476 (2009)
23. Shewokis, P.A.: Is the contextual interference effect generalizable to computer games? Perceptual and Motor Skills 84, 3–15 (1997)
24. Hatakenaka, M., Miyai, I., Mihara, M., Sakoda, S., Kubota, K.: Frontal regions involved in learning of motor skill–A functional NIRS study. Neuroimage 34, 109–116 (2007)
25. Leff, D.R., Elwell, C.E., Orihuela-Espina, F., Atallah, L., Delpy, D.T., Darzi, A.W., Yang, G.Z.: Changes in prefrontal cortical behaviour depend upon familiarity on a bimanual co-ordination task: An fNIRS study. NeuroImage 39, 805–813 (2008)
26. Gobel, E.W., Parrish, T.B., Reber, P.J.: Neural correlates of skill acquisition: decreased cortical activity during a serial interception sequence learning task. Neuroimage 58, 1150–1157 (2011)

Brain Activity Based Assessment (BABA)

Roy Stripling and Grace Chang

National Center for Research on Evaluation, Standards, and Student Testing (CRESST),
University of California, Los Angeles (UCLA)
10945 Le Conte Ave
Los Angeles, CA 90095, United States
stripling@cse.ucla.edu, gychang@ucla.edu

Abstract. Event-Related Potentials (ERP) are changes in brain activity detected using electroencephalographic (EEG) methods. One well-studied ERP is the P3b, which is generally elicited by asking participants to press a key when presented a target stimulus (e.g., "T") that is intermixed with a much more commonly presented non-target stimulus (e.g., "S"). We hypothesized that we could assess knowledge by asking participants to solve a problem then press a key when they see the correct answer in a series of (mostly wrong) answers. Early pilot testing (four participants) suggests that the P3b shows promise in this regard. In a math test, P3b responses were produced when shown correct, but not incorrect answers. In a foreign-language vocabulary test (matching picture to foreign word), P3b responses were not produced when shown correct answers prior to studying the words, but did produce P3b responses after studying. Some notable deviations in individual participants are discussed.

Keywords: Evoked Potential, Electroencephalogram, EEG, Knowledge Assessment.

1 Introduction

Assessing knowledge in learners, whether pencil-and-paper or though computer-based methods, typically involves an explicit question and answer process, where the answers are taken as evidence (for or against) learner knowledge. The scoring of such assessments may be straight forward (e.g., percent correct), may include individual test items that are weighted to account for differences in their a priori assessed difficulty, and/or may see item or overall scores adjusted based on statistical arguments that the learner was guessing on a given item (for full discussion of this approach, see the field of Item Response Theory [1],[2]). However, none of these approaches provides direct evidence that can distinguish correct answers that reflect true knowledge possessed by the learner from her guesses, or that can distinguish true misconceptions (wrong answers the learner believes to be right) from a simple lack of knowledge (wrong answers that are the result of guessing and/or that the learner knows she did not know). Empirical data suggests that the learner may, at least in some cases, lack introspective awareness of these differences, or at least, is not a reliable source of clarification[1],[2].

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 390–398, 2013.

Event-Related Potential (ERP) responses have been used as an additional source of direct evidence for possession of knowledge. ERPs are changes in brain activity that can be seen following presentations of stimuli to a person. They are measured using electroencephalographic (EEG) sensors, which detect small changes in the voltage potential on an individual's scalp. One particularly interesting ERP for this purpose is the P3b (also called the P3 and the P300). It is a transient positive shift in voltage observed from central EEG sensors, reaching its peak amplitude between 300-600 msec following presentations of "oddball" stimuli[3], [4]. It is independent of stimulus modality, but is typically stronger when the individual is consciously searching for the rare stimulus. Most P3b eliciting protocols expose participants to serial presentations of a non-target stimulus (e.g., the letter T), and ask the participant to perform a key press when they see the (much less common) target stimulus (e.g., the letter S). More complex presentations of non-targets (or distractors) also work, as long as the target is known and is relatively rare (10-20% of total stimuli presentations).

Most efforts exploiting ERP analysis in studies of learning and memory focus on gaining insight into the process of learning itself – i.e., what are the cognitive mechanisms of learning? However, a few efforts have sought the use of P3b detection as a method for knowledge assessment. These efforts have used the ERPs as evidence for word recognition, recognition of deviations of musical expectancy in experts versus novices, and for detecting "guilty knowledge" in criminal suspects.

Johnson, et al. [5] had participants study word lists, and then tested these participants for P3b elicitation during subsequent presentation of those words mixed with distractor words. They observed greater P3b amplitudes during presentation of studied words, which increased with the extent of studying permitted. Words that were correctly recognized elicited stronger P3b responses than study words that were recognized less consistently. Besson and Faita [6] studied musicians and non-musicians listening to musical phrases that were either selected from the classical repertoire or composed for the experiments. The musical phrases ended either congruously or with a musical violation. Musicians performed better than non-musicians in recognizing familiar musical phrases and classifying terminal violations. The ERPs (in this case an N400 ERP) to the end notes differed both in terms of amplitude and latency between musicians and nonmusicians, and as a function of participants' familiarity with the melodies and type of violation.

Detection of the P3b has been used (with some controversy) to determine if a criminal suspect posses knowledge of a crime that only the criminal or an investigator could know [7]. These suspects are typically shown a sequence of crime scene images. Most of the images in this sequence are not from the crime in question, but a few are. Detection of P3b ERPs in response to the images from the crime in question are taken as indicators of specific knowledge of the crime. If the suspect does not have a suitable explanation (e.g., they witnessed the crime, they investigated the crime, etc.), then these results are taken to connect them to the crime. The controversy with this approach is not whether it provides some useful information relevant to guilt or innocence; rather the controversy is related to the perfect accuracy rate claimed by its proponents [7].

These previous studies provide limited evidence that ERPs can be used to assess acquisition or possession of knowledge in some respect, but none provide a systematic exploration of the potential of ERPs in neuro-based assessments. What types of knowledge can be assessed? What form must the testing take to provide reliable valid, evidence of specific knowledge? What parameters can be manipulated without invalidating the approach? This paper provides a qualitative description of ongoing/preliminary work that is exploring whether ERPs, and in particular the P3b can be reliably used to assess possession of explicitly learned procedural and/or declarative knowledge. In the most common form of P3b eliciting experimental paradigms, P3b responses are elicited by rare target (visual or auditory) stimuli, presented as part of a series of non-target stimuli. Instead of instructing participants on what target stimulus they should search for, as is commonly done in P3 studies, we adapted this approach by presenting them with a problem and asking them to search for the correct solution in the set of answers that we presented to them serially. Our hypothesis was that by embedding the correct answer in a series of wrong answers, the correct answer (if recognized as such) would elicit a P3b response. Further, incorrect answers that the participant believes to be correct (reflecting misconceptions) will also elicit P3b responses, but both incorrect answers and correct answers that are not recognized by the participant will fail to elicit P3b responses.

2 Methods

All methods involving participants were approved by the University of California, Los Angeles (UCLA) Institutional Review Board. At the time of writing, four individuals (3 female, 1 male), age range 28-33, all fluent in English, have participated in this study.

2.1 Tasks

Each participant was asked to complete a series of 5 tasks. In each case, the participant was presented on screen instructions and told to press the space bar when they were ready to begin. They were also instructed to press the space bar when they saw the target stimulus (tasks 1 and 2) or the correct answer to the problem (tasks 3-5). Stimuli in all tasks were presented on screen for 500 msec. A single dot was displayed in the same location on screen for 2000 msec between each stimulus presentation. Participants were instructed that reaction time was not critical, but that they needed to press the space bar before the next stimulus appeared on the screen. They were also instructed to try not to blink while the stimulus was on the screen, but to blink a second or so after it went off screen. Their blinking pattern was surreptitiously observed during task 1 and feedback a reminder was provided if necessary.

Tasks 1 and 2 were replications of common P3 inducing protocols. In task one, a non-target stimulus (the letter “T”) was presented 90 times and a target stimulus (the letter “S”) was presented 10 times, randomly interspersed within the non-target

sequence, but not appearing within the first 5 presentations. Prior to beginning this task, participants were instructed to press the space bar when they saw the letter "S". In task two, participants were again instructed to press the space bar when they saw a new target stimulus (the letter "U"), which was presented a total of 10 times. But this time non-target stimuli (90 presentations total) were selected randomly from all of the other letters of the alphabet.

Tasks 3 through 5 were tests of explicit procedural or declarative knowledge. Task 3 asked participants to solve or simplify math equations. Twenty-two different problems were presented. When a problem was presented on the screen, participants were given as long as they needed to solve the problem, and then asked to press the space bar to initiate the sequential presentation of possible answers. To ensure participants focused on searching only for the correct answer, they were instructed that the answers might appear more than once, and that they should press the space bar every time they saw the correct answer. For the sequence of possible answers to each problem, one correct answer was presented (never in the first three presentations), and nine unique wrong answers were presented. Five of the wrong answers were repeated again (at random), for a total of 14 wrong answer presentations and only one correct answer presentation.

Task 4 and 5 tested participant recognition of ten common words in Pinyin (Chinese characters into Latin script). Task 4 tested their recognition of these words prior to being given the opportunity to study them, and task 5 tested them after studying them with provided flash cards. The words chosen were the Pinyin names of common animals (cat, dog, horse, pig, etc.). Each task used the same 10 words, but prompted the participant to identify them with different pictures of those animals. Likewise the flash cards included different pictures of the same animals used in tasks 4 and 5. Pre and post written tests were also given using different pictures to provide further evidence of whether the participant had prior knowledge of these words and/or had successfully learned them using the flash cards. As with task 3, the participant was presented a picture of the animal, asked to recall the Pinyin name of the animal, and to press the space bar to initiate the sequential presentation of possible answers. In this case, wrong answers were the names of the other nine animals. Answers were presented in random order. Five wrong answers were repeated again (at random), but the correct answer was not presented in the first three presentations and was presented only once.

2.2 Event Related Potential (ERP) Data Collection and Processing

Electroencephalographic (EEG) data were collected from each participant during all five tasks, using a B-Alert X10 EEG system (Advanced Brain Monitoring). The B-Alert X10 system records activity through nine sites (F3, F4, Fz, C3, C4, Cz, P3, P4, and POz) digitizing each at 256 samples per second. Event synching was achieved by processing bin files generated by the task presentation software. These files were processed in MATLAB in order to obtain the event (stimulus and response) and it's corresponding epoch and data-point. The epoch and data-point of each event was then stored in a common log file (CLF) that is processed with the .ebs file in the B-Alert

batch software. The resulting ERP outputs are time locked to the start of each stimulus presentation and is presented for 1 second (256 data-points).

Data were processed for all sites by staff at ABM who were blind to the conditions of the study, using standard methods for artifact detection and removal. Briefly, ERP waveforms that included artifact such as eyeblinks or excessive muscle activity were removed on a trial by trial basis using the B-Alert automated software. Additionally, trials with data points exceeding ±50μV were manually removed.

2.3 Quantitative/Qualitative Analysis

Quantitative analyses have not performed on the current preliminary dataset. Data collection from additional participants are ongoing and quantitative/statistical analyses will take place once the dataset is complete.

3 Results

Data were processed for the nine EEG channels recorded. However, P3b responses are attributed to central/posterior sources. For this reason, and because the data are preliminary, we report only descriptive results for POz. Cz and Fz displayed similar patterns across all subjects.

Figure 1 depicts the global average response across all participants as recorded from the POz location. Prominent P3b responses to target, but not non-target stimuli, are evident in trials from Task 1 and Task 2. Here we see peak amplitudes of approximately 20 μV ~450 msec after target stimulus presentation. Non-target stimuli peak amplitudes do not exceed 10 msec, and tend to peak closer to 300 msec post stimulus onset. Task 3 exhibits a weaker, but still evident P3b response to target stimuli. Target stimuli elicit an average response peaking at approximately 15 ~450 msec after stimulus onset. Non-target stimuli generate an average wave that is qualitatively similar to that observed in Task 2. In task 4, a P3b response is not evident to either target or non-target stimuli. Peak amplitude for either stimulus type is ~10 μV or lower and occurs ~350 msec after stimulus onset. Task 5 target stimuli, may exhibit a modest P3b response to target stimuli, but not to non-target stimuli. Target stimuli are associated with a peak amplitude of ~13 μV between 450 and 500 msec after presentation onset. Non-target stimuli are associated with a peak amplitude of less than 10 mV, with the peak occurring between 300 and 350 msec after presentation onset – qualitatively similar to non-target responses in tasks 2 and 3.

Participant variation from these averages are illustrated in figure 2. As can be seen in this figure, the first and third participants exhibit prominent P3b responses to target stimuli in task 3, while the second and fourth participant show no apparent P3b responses at all in this task. In task 5, participants three and four exhibit moderate to strong P3b responses to target stimuli. Participant 2 does not appear to produce a P3b response, but does show a very prominent negative response beginning around 500 msec after presentation onset that is selective for target stimuli. This participant also showed similar, but less intense pattern of response in Task 4 (which tested the same

stimuli but before the participants were allowed to study the words; data not shown). The first participant does not appear to produce a P3b response to target or non-target stimuli in this task.

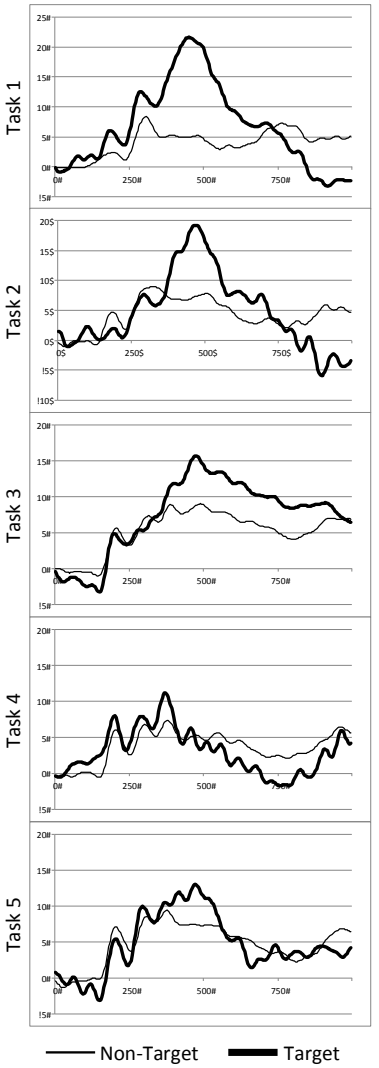


Fig. 1. Population ERP responses from each task following exposure to target and non-target stimuli. Tasks 1 and 2 replicate previous methods for inducing P3b responses. In task 3, participants were asked to solve math problems to determine the target stimuli. In tasks 4 and 5, participants were shown pictures of animals and told that the correct name for the animal in Pinyin (Chinese written using Latin characters) was their target. Participants were given the correct answers to study after task 4, but before task 5.

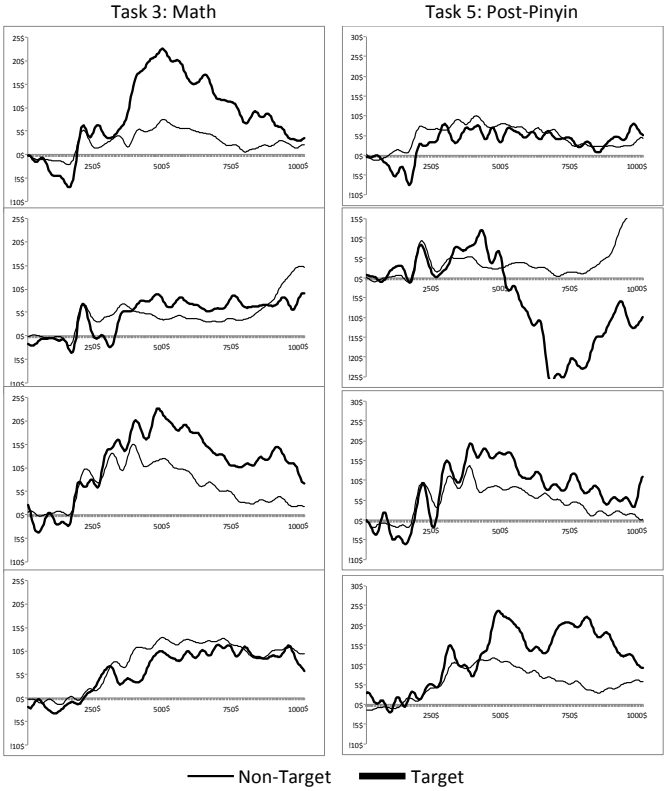


Fig. 2. Average ERP responses from individuals in Task 3 (Math) and Task 5 (Pinyin – after studying). Each row in this figure presents data from the same individual participant – averaged across their own target or non-target trials.

The first participant had no prior knowledge of Chinese, but was fluent in Korean. In discussing the tasks after the experiment, this participant indicated that while the written languages of Pinyin and Korean are very different, the spoken forms of some of the words are similar. This participant correctly answered 6 out of 10 of the words on the written pre-test. The second participant revealed some confusion about the anticipated solution format in the math task. In particular, this participant when confronted with problems that could be simplified but not solved (e.g., 2x+3x+ 5=__) assumed that the blank held a value of zero and solved the equation, rather than simplifying it. Despite this confusion, the participant selected correct answers in all but 3 of 22 problems. In addition, this participant revealed at the end of the experiment that they had prior exposure to Chinese, having taught English in rural China for a year. This participant correctly identified 4 of the 10 Chinese words in the pre-test. The third and fourth participants were given more explicit instructions with regard to solving versus simplifying problems. The third participant selected the correct answer in all but 4 problems, however the fourth participant selected 11 incorrect answers out of 22.

4 Discussion

Prior studies provide limited evidence that ERPs can be used to assess acquisition or possession of knowledge in some respect, but none provide a systematic exploration of the potential of ERPs in neuro-based assessments. This paper provides a qualitative description of ongoing/preliminary work that is exploring whether ERPs, and in particular the P3b can be reliably used to assess possession of explicitly learned procedural and/or declarative knowledge. In the most common form of P3b eliciting experimental paradigms, P3b responses are elicited by rare target (visual or auditory) stimuli, presented as part of a series of non-target stimuli. Instead of instructing participants on what target stimulus they should search for, we adapted this approach by presenting them with a problem and asking them to search for the correct solution in the set of answers that we presented to them serially. Our hypothesis was that by embedding the correct answer in a series of wrong answers, the correct answer (if recognized as such) would elicit a P3b response.

We began by establishing a baseline P3b response for each participant through tasks 1 and 2. The results of these tasks replicate prior results using similar if not the same paradigms. In addition, task 2 may have prepared the participant for our problem-solution variation by challenging them to find a specific target in a complex set of non-target stimuli. All four participants tested to date were able to discriminate the target from non-targets in tasks 1 and 2, and all generated robust P3b responses to targets, and not to non-targets in these tasks.

In tasks 3-5, we test our hypothesis by asking participants to determine what the target stimulus should be based on their knowledge of math procedures, or based on their knowledge of Chinese (written in Pinyin). We had assumed going in that our math problems were solvable by the population from which we would be recruiting, and that none of our population would have prior knowledge of Chinese. Instead, as described in the results section, we discovered that our math problems were sometimes confusing as written and our instructions on how to handle them too vague for some. This may have lead to a reduced P3b response to target stimuli, particularly for the second participant. Participant four did not produced an apparent P3b response to correct solutions in the math task, but their performance in that task suggests that this may be due to identifying incorrect solutions to the problem, in which case this lack of P3b would be consistent with our hypothesis.

We had also assumed that knowledge of Chinese language would be sparse in our participant population, but post-experiment discussions with our participants suggests otherwise. Participants one and two had partial knowledge of the words used in our tests, and both failed to generate a P3b responses in tests run before and after allowing them to study the words. The lack of any ERP response from participant one in this task does not support the hypothesis. The second participant did produce a late, negative ERP that was present in task 4 and more prominent in task 5 (post-studying). This is not consistent with our specific hypothesis that a P3b ERP should be elicited by recognized stimuli, but does suggests that other ERPs may also be a source of knowledge assessment. The particular ERP that reveals knowledge may differ based on cognitive strategies employed by the participant and/or may reflect some natural

individual variation that will have to be accounted for if ERPs are to be put to practical use in this regard.

Collectively, analysis of the data collected to date suggests that there may be potential for using ERPs, including the P3b, as a basis for knowledge assessment. The data also indicate that clear test items and unambiguous instructions are critical. In order to improve the quality and clarity of our test items, we plan to utilize math items taken from the National Assessment of Educational Progress (NAEP) database of test items. In addition, we will test alternative foreign languages to ensure that each participant is fully naïve in that task. And we will begin exploring history/civics test items also taken from the NAEP database to broaden the types of test items with which we test our hypothesis.

Acknowledgments. We thank Chris Berka and Veasna Tan from ABM for their assistance processing data for this project. This work was supported by funding from the U.S. Department of Education, award R305C080015.

References

1. de Ayala, R.J.: The theory and practice of item response theory, vol. xv. Guilford Press, New York (2009)
2. Lord, F.M.: Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Inc., Hillsdale (1980)
3. Comerchero, M.D., Polich, J.: P3a and P3b from typical auditory and visual stimuli. Clinical Neurophysiology 110(1), 24–30 (1999)
4. Polich, J.: Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology 118(10), 2128–2148 (2007)
5. Johnson, R., Pfefferbaum, A., Kopell, B.S.: P300 and Long-Term Memory: Latency Predicts Recognition Performance. Psychophysiology 22(5), 497–507 (1985)
6. Besson, M., Faïta, F.: An event-related potential (ERP) study of musical expectancy: Comparison of musicians with nonmusicians. Journal of Experimental Psychology: Human Perception and Performance 21(6), 1278–1296 (1995)
7. Farwell, L.A., Donchin, E.: The Truth Will Out: Interrogative Polygraphy ('Lie Detection') With Event-Related Brain Potentials. Psychophysiology 28(5), 531–547 (1991)

Part IV

Understanding and Modelling Cognition

Enhancing Intuitive Decision Making through Implicit Learning

Joseph Cohn, Peter Squire, Ivy Estabrooke, and Elizabeth O'Neill

Office of Naval Research, USA
Joseph.Cohn@navy.mil

Abstract. Today's military missions pose complex time-constrained challenges, such as detecting IED emplacements while in a moving vehicle or detecting anomalous civilian behaviors indicative of impending danger. These challenges are compounded by recent doctrinal requirements that require younger and less-experienced Warfighters to make ever-more complex decisions. Current understanding of decision making, which is based on concepts developed around theories of *analytic* decision making (Newell and Simon, 1972), cannot effectively address these new challenges since they are based on the notion of enabling experts to apply their expertise to addressing new problems. Yet, there are actually two types of recognized decision making processes, *analytical* and *intuitive*, which appear to be mediated by different processes or systems (Ross et al, 2004; Evans, 2008; Kahneman & Klein, 2009). *Analytical* decision making is mediated by processes that reflect a sequential, step-by-step, methodical, and time-consuming process. To be effective, *analytic* decision making appears to require domain expertise. In contrast, *intuitive* decision making relies upon a more holistic approach to processing information at a subconscious level (Luu et al, 2010). The thesis of this paper is that unlike *analytic* decision making, effective *intuitive* decision making does not require domain expertise but, rather, can be enhanced through training methods and technologies. This paper will explore ways in which the results from a range of studies at the behavioral, cognitive and neurophysiological levels can be leveraged to provide a comprehensive approach to understanding and enabling more effective *intuitive* decision-making for these non-experts.

Keywords: Cognitive Modeling, Perception, Emotion and Interaction, Intuition Decision Making, Implicit Learning.

1 Introduction

The traditional understanding of intuition suggests that it can guide the judgment process by assisting with the discovery of plausible solutions from which to choose (cf Bowers, et al. 1990). This characterization of intuition - and many others that follow from it (e.g Kahneman & Klein, 2009) - assumes a high level of familiarity with the information being detected. Yet a growing body of results ranging from the biological (mainly, neural) to the cognitive (Lieberman, 2000; Jung-Beeman et al., 2004;

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 401–409, 2013.

Luu et al 2010) suggests that pre-existing expertise, which requires years of practice to attain (Ericsson et al, 1993) may not be a key requirement for developing intuitive decision making processes. These studies suggest that intuitive decision making processes share some of the same underlying neural structures and cognitive processes as implicit learning (Frensch, 2003; Lieberman, 2000, 2007). By acquiring domain knowledge through implicit learning, one may be able to automatically strengthen, at the neural, cognitive and behavioral levels, the same capabilities that are needed for effective intuitive decision making (Figure 1), making intuition a strong candidate for enhancement through training.

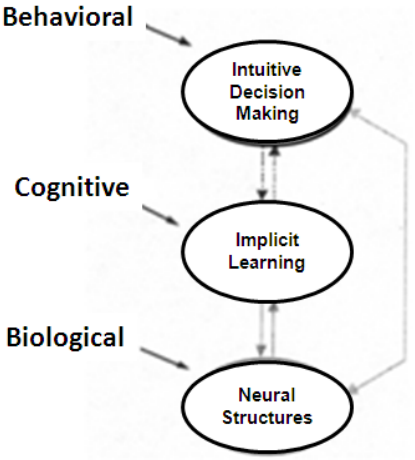


Fig. 1. Intuition relies on multiple layers of systems, from the biological to the cognitive to the behavioral. After Newell, 1993; Lieberman, 2000.

In order to develop these training capabilities, ***four challenges*** that are key to understanding and enhancing *intuitive* decision making must be addressed and understood: 1) Combining advances in measuring performance at multiple representation levels (e.g., neural, cognitive and behavioral) with advances in simulation–based paradigms for assessing decision making to understand the foundations of *intuitive* decision making; 2) Leveraging advances in cognitive modeling and machine learning techniques to represent individual *intuitive* decision making processes; 3) Developing an implicit learning based approach for enhancing *intuitive* decision making; and, 4) Combining these efforts, through scenario/simulation based training, to test and validate the hypothesis that implicit learning can enhance *intuitive* decision making for one or more operationally valid tasks. The remainder of this paper will discuss each of these challenges and possible solutions in greater detail.

1.1 Defining Intuition

Decision making is decomposed into two types or categories: *analytic* and *intuitive*. At the behavioral level *analytic* decision making is characterized by properties such

as deliberate and often lengthy periods of processing information, leading to a final result. At the cognitive level, *analytic* decision making seems to require intentional or goal-oriented information processing combined with a clear potential for being impacted by other cognitive processes – e.g. working memory. Finally, at the neural level, *analytic* decision making seems to be driven by a series of neural structures collectively acting as part of an (ad hoc) network. These structures include: Lateral Pre Frontal Cortex; Dorsomedial Pre Frontal Cortex; and Medial and Lateral Parietal Cortices and (Luu et al 2010; Lieberman, 2000, 2007; Bowers 1990). Perhaps most importantly, though, *analytic* decision making has shown itself to be accessible to a wide range of performance enhancement methodologies (Ericsson et al, 1993).

Conversely, *intuitive* decision making at the behavioral level is characterized by properties such as seemingly non-deliberate and fast operating information processing, seemingly at the pre conscious level. At the cognitive level, *intuitive* decision making seems to be cued by recognizable characteristics of the information being processed. At the neural level, *intuitive* decision making seems to organize brain networks for more advanced processing – acting as a 'coherence generator' for external information detected through sensory organs. Importantly, there have been only limited efforts focusing on enhancing *intuitive* decision making.

Because *analytic* decision making has proven to be more amenable to enhancement, the vast majority of efforts to improve overall decision making performance have focused on it. This bias towards *analytic* decision making belies the potential benefits to be gained by enhancing *intuitive* decision making. *Intuitive* decision making processes appear to provide a quick connection to the Limbic system ('gut responses') coupled with slower connections to frontal cortex and executive functions (Luu et al, 2010) potentially. As well, evidence suggests that intuition activates the formation of semantic networks in the brain. This means that *intuitive* decision making may actually set the stage for the detailed assessment of the benefits of taking those actions enabled by *analytic* decision making (Evans, 2008; Luu et al, 2010).

Figure 2a shows one way of envisioning this synergy between *intuitive* and *analytic* decision making, based notionally on the Human Information Processing notion of Parasuraman & Sheridan (2000) (Figure 2a). Early on the *intuitive* decision making system is activated, helping process key features of information while also priming the *analytic* decision making system. Later on, the *analytic* decision making system is activated, making sense of the information, enabling the decision maker to guide the process.

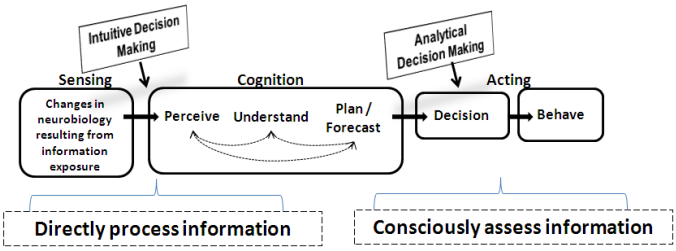


Fig. 2a. One view on how the two decision making systems may work together

Figure 2b provides a neural perspective on this synergy. When information requiring a decision and subsequent action is presented to an individual, initial features like contour and shape are registered by neural structures in the Temporal-Parietal-Occipital region (0 ms to ~250 ms). At approximately 250 to 300 ms other neural regions become activated, including those both in the limbic region, triggering the 'gut response' which is a hallmark of intuition, as well as those in cortical regions responsible for activating executive functions.

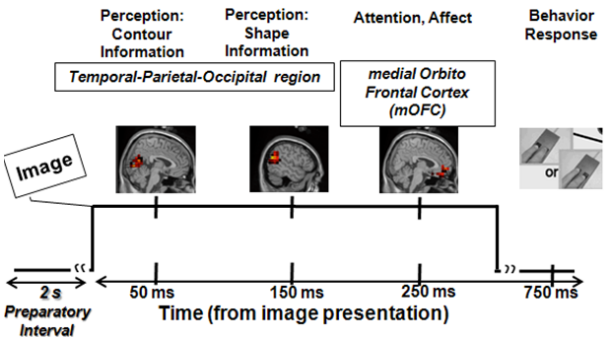


Fig. 2b. A neural level perspective on how the two decision making systems may work together. See text for details

1.2 Facilitating Intuition

The above discussion leads to two important points regarding enhancing human decision making. First, it suggests that we have a strong enough grasp of what intuition is, across different levels or representation, that we may consider it a ripe target for enhancement. Second, it suggests that by improving intuition we may streamline the decision making cycle. The critical question is how can we facilitate intuition?

Our starting point in addressing this question can be summed up by a quote from Reber, 1989: "To have an *intuitive* sense…is to have gone through an implicit learning experience." In order to develop one's *intuitive* capabilities, one must have moved through a type of learning known as implicit learning. In turn, this suggests that one may enhance one's *intuitive* capabilities through implicit learning. But what is implicit learning and why might it lead to enhanced *intuitive* decision making performance?

As Frensch & Runger (2003) define it, implicit learning is: "Learning complex information in an incidental manner, without awareness of what has been learned." Implicit learning emphasizes the role of associative learning mechanisms, coordinating action amongst different cognitive processes. Implicit learning exploits statistical dependencies in the environment, meaning that it is driven by certain kinds of features – or pattern structures-detected in information streams. Implicit learning leads to the generation of implicit knowledge as abstract representations (Seger, 1994) which provides the basis through which implicit knowledge can generalize to other contexts.

As Table 1 shows, the similarities between implicit learning and *intuitive* decision making are striking. Both processes seem to occur at a preconscious or unguided

level. Both processes appear to rely on recruiting different processes and or structures across the brain (Luu et al 2010). Both involve a level of pattern detection in the information stream being processed (Bowers, 1990). Lastly, both focus on transforming information into generalizable and actionable knowledge (Bowers, 1990). These similarities are equally striking when the neural structures underlying both processes are compared. Recent findings suggest that many of the neural structures that support intuition also support implicit learning (Luu et al 2010; Lieberman, 2000, 2007; Bowers 1990).

Table 1. Some similarities between implicit learning (Left) and intuitive decision making (Right)

Preconscious	
Coordinated Action	
• Implicit learning emphasizes the role of associative learning	• Intuition coordinates activity across the brain
Pattern Detection	
• Implicit learning exploits statistical dependencies in the environment	• Intuition requires perceiving coherence at a preconscious level
Generalization	
• Generates implicit knowledge as abstract representations for broader application	• Provides abstract 'hunches' about the nature of the pattern in question

Together, these findings suggest that from a neural, cognitive and behavioral perspective we may be able to facilitate intuition through implicit learning. The question then becomes how best to do this. We propose a four step process that includes:

- Characterizing *intuitive* decision making and implicit learning across neural, cognitive and behavioral levels of representation
- Representing *intuitive* decision making through cognitive models in order to guide implicit learning techniques.
- Applying scenario based training techniques to develop implicit learning approaches that enhance *intuitive* decision making.
- Testing the hypothesis that implicit learning facilitates *intuitive* decision making.

1.3 Techniques for Characterizing Intuition

Traditionally, *intuitive* decision making has been studied at the behavioral level only, relying on simple reaction-time measures to infer when *intuitive* decision making has occurred (Hodgkinson et al, 2008). Recent developments in the cognitive neurosciences suggest that it is possible to characterize intuition across multiple levels of representation, thereby gaining deeper insight into how intuition works. For example, Lieberman et al (2007) showed that the two decision making systems are actually

driven by two separate networks of brain areas while Luu et al (2010) demonstrated that it was possible to directly correlate decision making behaviors with neural markers derived from activity in these two systems to determine when intuition occurred and when it did not.

There are a wide range of technologies that can be used to detect intuition. These technologies can be categorized in terms of 'Data Source' 'Measurement Time', and 'Data Channels'. Figure 3 provides a representation of some common types of detection technologies in terms of these three parameters. On the 'Data Source' axis, the data sources that may be accessed to characterize intuition range from subcellular processes and individual nerve cell action, to measured behavior outcomes. On the 'Measurement Time' scale, the different time scales underlying the processes represented by each of these data sources are shown. On the 'Data Channels' axis, the number of measurable Data Source 'units' is represented. For example, the potential number of ion channels that could be measured, limited by the size of probes, or the number of behavior responses, limited by the number of metrics that can be associated with a given action.

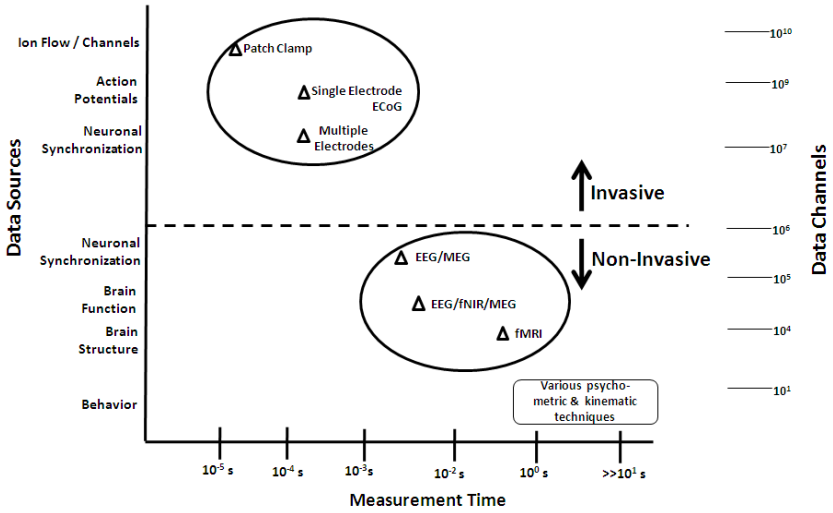


Fig. 3. Detection technologies

1.4 Techniques for Modeling Intuition

In order to make the characterization of intuition accessible to new training technologies an "executable" representation of these data must be developed. This requires both new approaches to decoding the performance data and to representing it as a model. Over the past several years, various machine learning techniques have been developed that help organize large, multi-scale sets of time series data into information classifiers. These multivariate decoding routines (Mitchell et al 2004) have the ability to take into account the full spatial pattern of brain activity, cognitive measures

and behavioral outcomes and appear to be transferrable to other, never-before encountered individuals, with little reduction in accuracy (Shinkareva et al 2008). In practice, it is expected that the initial classification routines will require a wide range of data sets and types, encompassing biological, cognitive and behavioral.

These classification approaches provide the rectified data necessary for building models of human performance. One approach that continues to gain momentum is to take existing cognitive models and link them to neural data. For example one of the better known cognitive modeling approaches is ACT-R (Anderson, 1996). In its executable form, the timing and sequencing of ACT-R's model components is based on observed behaviors, and the output is typically timing and accuracy predictions. Recently, studies performed by Anderson et al (2008) have demonstrated which neural regions correspond to which elements of their modules and buffers, opening up the possibility for a direct link between neural data and a proven cognitive modeling approach. Other approaches focus on developing *neurocognitive architectures* that are specifically tailored to fuse data captured from different sources to create generative hybrid models (see Figure 4). These approaches blend top-down and bottom-up approaches to innovatively combine context with various types of cognitive and neural measures to model overall user performance. Using machine learning and artificial intelligence routines, these approaches can adapt their models using past behaviors, specific actions taken and outcomes realized.

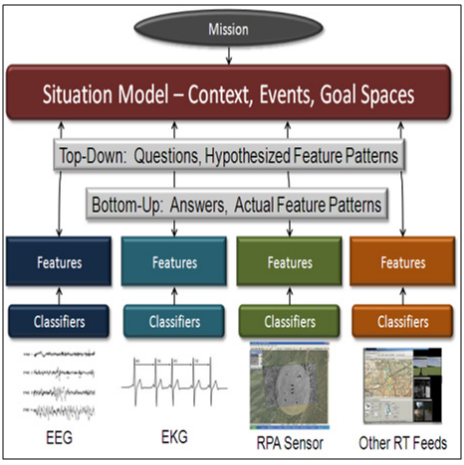


Fig. 4. An example neurocognitive architecture. Shown are both the *top-down* aspects, like hypothesis generation about patterns in collected data as well as *bottom-up* processes, like the data fusion and classification (with permission from Dr Webb Stacy, Aptima).

1.5 Scenario-Based Training

The classical understanding of intuition is that it requires a high degree of domain expertise. By some estimates, achieving such expertise may require up to ten years of intense exposure to any number of a wide range of practical 'training' exercises, which is well outside the training cycle in which effective Marine decision makers are developed. As envisioned here, *intuitive* decision making develops as a result of the

'strengthening' of connections with specific structures in the brain, like the basal ganglia, combined with the development of specific types of targeted training, collectively known as implicit learning. In practice, implicit learning is experiential and interactive, instead of didactic and classroom based. Therefore, it seems reasonable to focus on training technologies like virtual environments or serious games to provide the "experiential" component, using models of an individual's *intuitive* processes to modify the "interactive" component. The overarching training methodology to be employed will be Scenario Based Training (SBT), which emphasizes embedding training approaches within an evolving and dynamic scenario rather than delivering it through a series of static lessons (Oser et al 1999).

1.6 Measuring Success

There are two possible approaches for measuring success in this kind of effort. The first is to demonstrate that the neural structures that are active during implicit learning are also active during intuition; that in the absence of implicit learning there are different / distinct patterns of neural activity during an *intuitive* decision making task; and that in control tasks in which neither implicit learning was provided or *intuitive* decision making required, these structures are minimally active. This approach will essentially compare measures of neural activity across different task conditions.

The second is to show that, under those conditions in which implicit learning was provided and *intuitive* decision making was present, there is a significant improvement in decision making compared to other conditions as represented for instance by a shift in the form of receiver operator characteristic curves.

2 Summary

This paper proposes a new approach for enhancing *intuitive* decision making in novices, outlining four areas to address in order to develop training technologies for *intuitive* decision making. First, the nature of *intuitive* decision making must be characterized, at the neural, cognitive and behavioral levels. Second, these characterizations must be integrated into a single model that accurately represents these characterizations providing the foundation for developing training technologies. Third, the resultant model must be implemented into a training technology that demonstrably enhances an individual Warfighters' *intuitive* decision making capabilities. Finally, the effectiveness of this approach must be determined through a range of assessment techniques.

References

1. Anderson, J.R.: ACT: A simple theory of complex cognition. American Psychologist 51, 355–365 (1996)
2. Anderson, J.R., Carter, C.S., Fincham, J.M., Qin, Y., Ravizza, S.M., Rosenberg-Lee, M.: Using fMRI to Test Models of Complex Cognition. Cognitive Science 32, 1323–1348 (2008)

3. Bowers, K.S., Regehr, G., Balthazard, C.G., Parker, K.: Intuition in the context of discovery. Cog. Psyc. 22, 72–110 (1990)
4. Ericsson, K.A., Krampe, R.T., Tesch-Romer, C.: The role of deliberate practice in the acquisition of expert performance. Psychological Review 700, 379–384 (1993)
5. Evans, J.: Dual-processing accounts of reasoning, judgment, and social cognition. Ann. Rev. Psyc. 59, 255–278 (2008)
6. French, P.A., Runger, D.: Implicit Learning. Current Directions in Psychological Science 12, 13–18 (2003)
7. Hodgkinson, G., Langan-Fox, J., Sadler-Smith, E.: Intuition: A fundamental bridging construct in the behavioral sciences. British Journal of Psychology 99(1), 1–27 (2008)
8. Jung-Beeman, M., Bowden, E.M., Haberman, J., Frymiare, J.L., Arambel-Liu, S., Greenblatt, R., et al.: Neural activity when people solve verbal problems with insight. PLoS Biology 2, 500–510 (2004)
9. Kahneman, D., Klein, G.: Conditions for intuitive expertise: A failure to disagree. Am. Psyc. 64(6), 515–526 (2009)
10. Lieberman, M.D.: Intuition: A social cognitive neuroscience approach. Psychological Bulletin 126(1), 109–137 (2000)
11. Lieberman, M.D.: Social cognitive neuroscience: A review of core processes. Annual Review of Psychology 58, 259–289 (2007)
12. Luu, P., Geyer, A., Wheeler, T., Campbell, G., Tucker, D., Cohn, J.: The Neural Dynamics and Temporal Course of Intuitive Decisions. Public Library of Science (2010) (in Press)
13. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to Decode Cognitive States from Brain Images. Machine Learning 57, 145–175 (2004)
14. Newell, A., Simon, H.A.: Human problem solving. Prentice-Hall, Englewood Cliffs (1972)
15. Oser, R.L., Cannon-Bowers, J.A., Salas, E., Dwyer, D.J.: Enhancing human performance in technology-rich environments: Guidelines for scenario based training. In: Salas, E. (ed.) Human Technology Interaction in Complex Systems, vol. 9, pp. 175–202. JAI Press (1999)
16. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 30, 286–297 (2000)
17. Ross, K., Klein, G., Thunholm, P., Schmitt, J., Baxter, H.C.: The Recognition-Primed Decision Model Mil Rev., p. 6–10 (2004)
18. Reber, A.S.: Implicit learning and tacit knowledge. Journal of Experimental Psychology: General 118, 219–235 (1989)
19. Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., Just, M.A.: Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. PLoS ONE 3, e1394 (2008)

Measuring Engagement to Stimulate Critical Thinking

Patricia J. Donohue, Tawnya Gray, and Dominic Lamboy

San Francisco State University, Instructional Technologies
San Francisco, California, USA
pdonohue@sfsu.edu

Abstract. This research is a theoretical study of game-augmented instruction for learning and playing mathematics challenges. We wanted to extend our work with a unique Studio-Based Learning (SBL) model for peer-critiques of project designs. SBL had been used successfully in 15 universities as an approach for helping undergraduate computer science students improve their programming skills and code reviews. We piloted the model in a 9th-grade spatial studies class with some success in teaching freshmen how to critique their work and participate in peer reviews across teams. From those experiences we developed a framework for an interactive mobile application of the studio experience. Research with a group of student athletes revealed that before mobile development, we needed to consider the constraints of learner characteristics on the mobile environment. This study sets out the design for a pilot test of our finding that learning style may drive game features for instruction.

Keywords: Mobile Learning, Mathematics, Physiological Measurement, Engagement, Physical Cognition, Game Theory.

1 Background

Research has made great strides in understanding the potential of virtual games to engage learners and advance their learning as effectively or beyond the traditional lecture (Magana, 2009). Understanding the potential that game-based features can bring to instruction is insufficient without first understanding the potential that learners bring to the game. A classic issue for any instructional systems design solution is how to match the pedagogical approach to the learners and the learning context (Dick, Carey and Carey, 2007). We considered the problem of trying to generalize instructional game approaches for all learners, and asked the question: Do specific types of learners require different types of game features for the instruction to be effective? While it may seem an obvious yes, research has more often looked at how to apply known gaming approaches to improve instruction without consideration for the complexity in the learning population. Before we could evaluate the students-to-mobile system for effectiveness in facilitating student critiquing, our research showed we needed to ask more questions about the mobile environment and the students engaged with it. We decided to start with a known instructional problem for a known low achieving student population to investigate how a game-based approach might engage

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 410–417, 2013.

these students in mathematics problem solving. We selected 9th-grade student athletes with low achievement in mathematics as our study population. We then considered which game features might help the learner gain deeper content understanding, and decided on a study of differences in student engagement in a mobile mathematics game, controlled for student learning characteristics.

Our original research was funded on work from a CPATH (Computing Pathways) grant awarded by the National Science Foundation (NSF) to a university collaborative of principal investigators from Auburn University, Washington State University and the University of Hawaii at Manoa. A first CPATH grant (CPATH I) was funded from 2007-2010, and involved the development and testing of a unique Studio-Based Learning model for teaching undergraduates in computer science. A second CPATH grant (CPATH II), from 2010-2012, focused on expanding dissemination of the SBL model to new student audiences. Both grants reported increased engagement in programming tasks and slight, steady improvements in achievement (CPATH II Annual Report, 2011). The foundation of the model was the *Design Crit* (*Crit*), where students participated in peer critiques following a protocol for conducting code reviews. The *Crit* idea was generated from the master-apprentice relationships in the architectural studios of the 19th-Century. The Studio allowed for ongoing critique of students' work by masters and peers at any time in the design process. The SBL classroom is designed as a production studio where students can work, meet in teams, and seek instructor feedback. The NSF offered a Research Experiences for Teachers (RET) supplemental grant to fund expansion of existing CPATH work to K-12. Our research group applied and received a RET grant to pilot the SBL model in the 9th-grade Spatial Studies class at New Technology High School in Napa, California for the academic year 2010-2011.

The two teachers who had been teaching the geometry and digital media arts classes had just been assigned to integrate their classes into a new block class, called Spatial Studies. The grant provided opportunity to write the curriculum and it was developed entirely around a customized SBL model fit to New Tech High's project-based curricula (Donohue, 2011). Replicating the core SBL model from computer science, the Spatial Studies class focused on teaching students to think critically and conduct peer reviews on their geometry and digital designs. We found that students preferred the SBL model over the project-based model at New Tech because it provided them more structure – they "knew what to do next" – and, it gave them more resources from peers in the design reviews. Students reported that they enjoyed learning geometry on the computer as well as on paper and felt that they learned many more digital skills Photoshop, Excel, Illustrator, and GeoSketchpad. Interviews with students at the beginning and end of the class showed that they formulated their learning more precisely and used mathematics terminology more accurately by the end of course. While that would be a natural outgrowth from one year of learning, students attributed their improved understanding to their team critiques and peer-experts, a phenomenon also reported by the teachers.

Our current work in Instructional Technologies at San Francisco State University has focused on the effectiveness of mobile learning and the idea of bringing SBL to a mobile device provided a spark for a small group of faculty and students who

accepted the challenge to study a match of learner characteristics to mobile learning games. Our research question asked: Could we engage students in learning geometry by using a mobile game-based approach? If the answer was *Yes*, then we knew we could design a mobile application to facilitate students' critical thinking in geometry.

2 Theoretical Foundations of the Study

This research study focused on the problem of how to engage and help low achieving students learn geometry. We know from the research that computer games and simulations (Regan, et al., 2005; Fairclough, 2009; Magana, Brophy, and Bodner, 2010) can improve student understanding and build self-confidence in learning domain concepts. This was encouraging for a mobile game solution to teach geometry, especially for middle and high school students. We turned our focus to the match of learner to game.

Our study investigated the proposed game's ability to engage low performing students in mathematics. The population consists of sixth and seventh grade students in school sports. We postulated that this group of potentially low achievers in mathematics would be more likely to be engaged in geometry if they could participate in it physically. Our theoretical parameters involved research at the nexus of four fields of inquiry with potential impact on the study's outcomes:

1. The historical use of computing to teach mathematics
2. The success of virtual manipulatives to assist mathematics cognition
3. The principles of Mayer's Multimedia Learning theory
4. The implications of Gardner's Multiple Intelligences

The result was a game pilot that appeals to student athletes' heightened bodily-kinesthetic and visual-spatial intelligences. We developed two mobile applications of a basketball competition for testing: one for a mobile phone and one for Microsoft Kinect 360. The game incorporates findings from virtual manipulatives and multimedia learning theory to shape the environment for greatest effect.

2.1 Historical Use of Computers

James Kaput and Jeremy Roschelle (1998) proposed in their chapter review, *The mathematics of change and variation from a new perspective: New content, new context,* that a Dual Challenge existed with the growth of technology that required teaching "more math to more people" (section *Dual Challenges: Much more mathematics for many more people*, par. 1). They pointed out that, by the turn of the century, teaching mathematics had become increasingly abstract and complex in the face of increasing student diversity and social cost. While pointing out the advantages that new technologies offered, they concluded their review with the question "Can these new possibilities transform our notion of a core mathematics curriculum for all learners?" Their emphasis on "all learners" alluded early to the inability of mass solutions to meet

individual needs. More recent work in artificial intelligence that allows for individualized instruction with solutions such as the Cognitive Tutor (Koedinger, 1998) or adaptive testing have run into the same constraints of increased cost, learner diversity, and complexity of content. Our challenge to match the learner to the system would not be a trivial question.

2.2 Virtual Manipulatives

Physical manipulatives have been successful teaching tools in mathematics since their introduction into schools in 1989. Various manipulatives such as base 10 blocks, colored chips, interlocking cubes, and geo-boards proved their worth in helping students conceptualize abstract concepts. Manipulatives allow students to make abstractions meaningful. They facilitate learning by making relationships between ideas explicit using visual, tactile and kinesthetic experiences (Hunt, Nipper and Nash, 2011).

Virtual manipulatives (VM) have shown new potential to engage learners by offering unique characteristics that go beyond the capabilities of physical manipulatives (Moyer-Packenham, Salking and Bolyard, 2008). While virtual affordances can enhance the user's experience and understanding of a mathematics concept, they can also detract or disrupt attention and perception. VM can have drawbacks if not designed well. The visuals can be distracting or disorienting in use, but the authors note that VM was most effectively used in tests with third-grade students when applied in the middle or core part of a lesson: "It was during these activities (investigation and skill solidification) that teachers reported the engagement of the students with the virtual manipulatives" (p.214).

In looking at multimedia principles applied to virtual manipulatives, Packenham et al. (2008) point out that "Dual Coding Theory (Clark & Pavio, 1991) and Multimedia Principles (Mayer & Anderson, 1992) support the notion that when learners are presented with visual and verbal codes, the effects of multimedia instruction and students' recall of information are increased" (p.214). The findings of their study showed that virtual manipulatives "were central to the mathematics learning and content development and were often used in combination with physical manipulatives" (p.215). Our game would need to build on the success of VM in developing content learning.

2.3 Multimedia Learning Theory

Richard Mayer's (2001) Multimedia Learning Theory states that instructional messages should be developed in light of how the human mind works. Mayer's research shows how words and pictures are qualitatively different yet complement each other and that human understanding occurs when learners are able to integrate visual and verbal representations. By building connections between words and pictures, learners are able to create a deeper understanding than from words or pictures alone.

Mayer (2001) bases his cognitive theory of multimedia learning on three main assumptions: 1) Dual Channel - states that humans possess separate channels for visual and auditory information; 2) Limited Capacity - states that humans are limited in the amount of information they can process at one time; and 3) Active Processing - states that humans have meaningful and transferable learning experiences when they engage

in active learning as defined by "attending to relevant incoming information, organizing selected information into coherent mental representations, and integrating mental representations with other knowledge." Our game interface would need to make use of multimedia design principles.

2.4 Multiple Intelligences

Howard Gardner introduced the theory of Multiple Intelligences with his book *Frames of Mind* in 1983. Multiple intelligences theory challenges our traditional notion of intelligence. He argues that multiple intelligences deny the application of a universal or mass approach to measure intelligence, such as the IQ (Intelligence Quotient) test. This suggests that current approaches to instructional development using game theory and gamification approaches might miss the critical determining factor of individual differences. One advantage of games is their potential for customization or personalization by the user. However, the ability of the user (learner) to select or dress the player in the game to suit his or her preferences does not address the need alluded to here for learners to choose a type of game that fits his or her learning approach.

Our selected learners for intervention are student athletes involved in school sports. Gardner (1983) explains intelligence as raw biological potential to process information and problem solve. The two intelligences important to this study are the bodily-kinesthetic and visual-spatial intelligences.

Bodily-kinesthetic intelligence has been defined as "the ability to problem solve or fashion products using one's whole body, or parts of the body" (Gardner, 1993). Bodily-kinesthetic learners process information through the sensations they feel in their bodies and tend to learn through movement and touch. Individuals with this intelligence prefer to communicate information by demonstration and modeling. These learners include athletes, dancers, actors and surgeons.

The visual-spatial intelligence has been defined by Gardner (1993) as, "the ability to form a mental model of a spatial world and to be able to maneuver and operate using that model." Individuals with high visual-spatial intelligence tend to think in pictures and are able to learn readily from visual presentations. Our most surprising finding alerted us that our game would likely be more effective if we could meet our student athletes' learning preferences.

2.5 Ways That Games Engage

We chose to build a mobile application based on the principles and lessons of multimedia learning, virtual manipulatives, and multiple intelligences noted above. We know from the research that gamification of instruction offers numerous ways to engage learners in content: rewards for achievement, instant feedback, enhancement of attention, a state of uncertainty that triggers Dopamine for heightened enjoyment, and engagement by playing with other people (Chatfield, 2012; Gee, 2005; Camerer, 2003). Given these challenges for design and development of the platform, we designed a study as outlined in the methods section that follows.

3 Methods

We chose to design an Augmented Basketball Challenge. Our interests for the pilot were to explore three areas of investigation: 1) the role of visual-spatial and bodily-kinesthetic intelligences on learning with the game, 2) the attraction of a mobile (or virtual) game to engage low-performing students in learning, and 3) the ability of physically augmented cognition to impact students' conceptual thinking in geometry.

Our first prototype, the Augmented Basketball Challenge, places young players in friendly competition to demonstrate their understanding of triangles, angles, parallel and perpendicular lines. More than a classroom manipulative; more than a simulation; the Augmented Basketball Challenge uses a virtual 3D competition to stimulate student understanding of mathematics by physically manipulating a visual basketball in live play. The pilot test will explore the physical-cognitive link during game play. We know from Purdue's worldwide research on Nanohub.org (Magana, Brophy, and Bodner, 2010) that simulations can improve student understanding and build self-confidence in learning domain concepts. We know, as Howard Gardner (1993) suggests, that students' spatial and bodily-kinesthetic intelligences act as cognitive aids to learning when body, mind, and game converge. To gain deeper insight into the potential effects of these principles on mathematics learning, we are conducting a mixed methods study of the game's implementation with middle school students.

We have chosen the methods employed by Regan, Mandry, Kori, Inkpen and Calvert (2005) in their study using questionnaires, interviews, video coding of observations, and Galvanic Skin Response (GSR) to measure the user experience with entertainment technologies. While their study used a hybrid game system to analyze the differences between computer systems, we will collect the same type of data from four student groups in a 2x2 design: student athletes with low mathematics scores on the 8th grade high stakes testing, student athletes with average to above average mathematics scores, student non-athletes (scoring low on bodily-kinesthetic and visual-spatial intelligences testing) with low mathematics scores, and student non-athletes with average to above average mathematics scores. Each group will participate (within-group) in the Augmented Basketball Challenge on the mobile application and then one round on the Kinect 360. Videos during game play will capture gestures, facial expression, and audio. A likert-based questionnaire before participation will capture student perceptions and attitudes towards mathematics and geometry in particular, experience with video or online games, and any experience with educational games. Post group interviews will be taped and conducted immediately after participation. The interviews will collect information on students' perceptions of the game and their experience, attitudes and perceptions on the geometry challenges, their preferences for learning with mobile devices and their observations of the game experience.

3.1 The Game Design

The game covers geometry basics from the 9th-grade mathematics standards on identifying triangles, angles, parallel and perpendicular lines. The game gives students opportunity to practice the standards and assess their learning. They also have the

chance to compete with other students and engage in the social aspects of the game. The mobile application model presents live basketball footage that demonstrates a standard. The video is enhanced with graphics to add demonstration of the concept. For example, to demonstrate a right angle, a player may pass the ball from one player to another player and then a third player forming a right triangle. The result is shown with arrows on screen. Student mathematics' assessment will occur through competition, either with the computer or against each other. Game competition will be timed and continuous. The time is shown as a shot clock on a basketball court. The game presents students with a series of mathematics terms, given one at a time, in random order. The player must perform the concept of the term presented and shoot. Every correct calculation scores a point. The highest point wins. The game will be personalized for players with their picture added to a player's scorecard.

iPad Simulation. The student uses fingers to swipe the motion of the ball in a trajectory. After completing an angle, the student shoots the ball by swiping it towards the basket. If the basket is made, the student's calculation was correct. If the basket is missed, the calculation was incorrect. Lines and Arrows demonstrate where the ball has been passed (See Fig. 1). Onscreen colors signal correctness of actions and rewards are used to encourage play and challenge the learner.

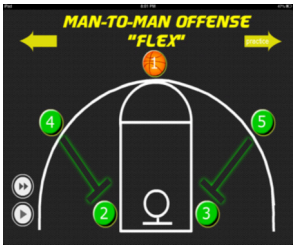


Fig. 1. Screen shot of test iPad application for teaching basketball plays, showing ball trajectories (green) and touch controls at lower left. Developed by Tawnya Gray.

XBOX 360 Kinect Simulation. The game operates as in the iPad simulation; however, instead of a swipe of the finger, the student simulates a passing motion of the ball in the trajectory desired, for the player onscreen to catch it. The student is always the player with the ball. When the student *feels* the correct answer they will shoot the ball (the same way as if they were on a real basketball court). Onscreen, the game shows the player shooting the ball. If the student's calculation is correct, the ball makes the basket; if incorrect, the student misses the basket.

4 Implications of the Study

In 2011 there were 34,024 student-athletes participating in the sport of basketball for NCAA (National Collegiate Athletic Association) affiliated institutions of higher learning. Most of these students enter four-year institutions as freshmen and are required to learn a significant amount of basketball tactical plays in order to compete. College basketball tends to have a high attrition rate of freshmen student-athletes who must successfully make the transition from high school to college. We looked at this population of students who are challenged on at least three fronts. They must successfully transition from high school to college academic standards, athletic standards, and to a higher-level basketball game. These students were often the low achievers in

mathematics and other academic disciplines. If we can teach these athletes mathematics using a physical basketball game framework in high school, we might be able to help collegiate athletes before they reach their freshman year learning trauma.

References

1. CPATH II: Broadening Studio-Based Learning in Computer Education. Annual Report: 01.2010-12.2010 (February 6, 2011)
2. Camerer, C.F.: Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press (2003)
3. Chatfield, T.: Seven Ways Games Engage, https://www.youtube.com/watch?v=KyamsZXXF2w (accessed January 25, 2013)
4. Clark, J.M., Paivio, A.: Dual coding theory and education. Psychology Review 3(3), 149–210 (1991)
5. Donohue, P.J.: A Studio-Based Learning model for K-14 computational thinking in STEM content, A report prepared for the National Science Foundation Site Review, April 14-15. University of Hawai'i at Manoa & Hilo (2011)
6. Gardner, H.: Frames of Mind. Basic Books, Inc., New York (1983)
7. Gardner, H.: Multiple Intelligences: The Theory in Practice. BasicBooks, New York (1993)
8. Gee, J.P.: Learning by Design: good video games as learning machines. E-Learning 2(1), 5–16 (2005)
9. Mayer, R.: Multi-media Learning. Cambridge University Press, Cambridge (2001)
10. Mayer, R.E., Anderson, R.B.: The instructive animation: Helping students build connections between words and pictures in multimedia learning. Journal of Educational Psychology 84(4), 444–452 (1992)
11. Koedinger, K.R.: Intelligent Cognitive Tutors as Modeling Tool and Instructional Model. In: A Position Paper for the NCTM Standards 2000 Technology Conference, June 5-6. Carnegie Mellon University (1998)
12. Moyer-Packenham, P.S., Salkind, G., Bolyard, J.J.: Virtual manipulatives used by K-8 teachers for mathematics instruction: Considering mathematical, cognitive, and pedagogical fidelity. Contemporary Issues in Technology and Teacher Education 8(3), 202–218 (2008)
13. Regan, L., Mandryk, K.M., Calvert, T.W.: Using psychophysiological techniques to measure user experience with entertainment technologies. Journal of Behaviour & Information Technology (2005), http://www.reganmandryk.com/pubs/mandryk_bit_preprint.pdf (accessed February19, 2013)

Human Dimension in Cyber Operations Research and Development Priorities

Chris Forsythe[1], Austin Silva[1], Susan Stevens-Adams[1], Jeffrey Bradshaw[2]

[1] Sandia National Laboratories, Albuquerque, NM, USA,
[2] Institute for Human and Machine Cognition, Pensacola, FL, USA
{jcforsy,aussilv,smsteve}@sandia.gov,
jbradshaw@ihmc.us

Abstract. Within cyber security, the human element represents one of the greatest untapped opportunities for increasing the effectiveness of network defenses. However, there has been little research to understand the human dimension in cyber operations. To better understand the needs and priorities for research and development to address these issues, a workshop was conducted August 28-29, 2012 in Washington DC. A synthesis was developed that captured the key issues and associated research questions.

Research and development needs were identified that fell into three parallel paths: (1) human factors analysis and scientific studies to establish foundational knowledge concerning factors underlying the performance of cyber defenders; (2) development of models that capture key processes that mediate interactions between defenders, users, adversaries and the public; and (3) development of a multi-purpose test environment for conducting controlled experiments that enables systems and human performance measurement.

Keywords: Applications of Augmented Cognition, Cyber Security, Research, Human Factors, Cognitive Modeling.

1 Introduction

Within cyber security, the human element represents one of the greatest untapped opportunities for increasing the effectiveness of network defenses. However, there has been little research to understand the human dimension in cyber operations. To better understand the needs and priorities for research and development to address these issues, a workshop was conducted August 28-29, 2012 in Washington DC. The findings of the workshop are summarized in this report.

The workshop brought together operational, scientific and programmatic perspectives, with the objective to converge upon a prioritized list of key research questions. While the human dimension encompasses defenders, attackers and users, for the current workshop, emphasis was focused only upon defenders. A range of topics were considered that contribute to increasing the effectiveness of cyber defenders, while minimizing the impact on users.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 418–422, 2013.

The workshop consisted of a series of focused discussions. The scope encompassed all areas impacting the effectiveness of cyber defenders in accomplishing their mission. This included (1) understanding the cognitive processes, (2) application of technology to support and enhance cognitive performance, (3) work processes/ environment and other factors that mediate performance, (4) collaboration and teamwork, (5) education and training, (6) organizational and cultural factors, and (7) personnel selection and retention.

2 What Are the Key Research Questions?

Research questions were identified that fell into several somewhat overlapping categories. The following sections discuss the core issues underlying these categories.

2.1 Measurement and Metrics

For the most part, there currently exists no quantitative basis for assessing the performance of cyber defenders, whether at the individual, team, group or organizational levels. Furthermore, while various resources are available for generating simulated cyber events and observing the behavior and performance of cyber defenders, without underlying science regarding the human dimension within cyber and the associated phenomenology, there is little basis for making decisions concerning the specific nature of exercises, who participates and how performance is evaluated.

2.2 Human Performance of Cyber Defenders

From a scientific perspective, there is very little known about cyber analysts. As a basis for scientific study, there is need for analysis to understand the jobs filled by cyber analysts, and particularly, the associated cognitive processes that mediate their performance.

2.3 Understanding the Adversary

It may be generally assumed that there is benefit for the cyber defender to have an understanding of their adversary. However, there is need for research to understand what types of knowledge is beneficial and how that knowledge may be effectively put into use.

2.4 Selection and Training of Cyber Defenders

Currently, there is little known about what attributes prepare an individual to become an effective cyber defender. There is little understanding of what skills, knowledge and abilities need to be addressed through selection and training. Likewise, within the course of training, there is need for research to scientifically establish the appropriate

measures for assessing performance, as well as approaches for effectively diagnosing and intervening to maximize training effectiveness.

2.5 Intersection between Humans and Technology

Building upon a better understanding of cyber defenders, questions arise concerning the balance between humans and technology, and how technology may be employed to augment the performance of individuals and teams.

These questions generally fall into two related areas. First, which cognitive processes operating at either the individual or team level should technology be used to augment and what mechanisms might be employed to do so. Second, what technologies would be most beneficial to the cyber defender (e.g. data mining, anomaly detection) and for these technologies, how should they be implemented?

2.6 Teamwork and Collaboration

Cyber defense often requires the effective coordination of teams. However, there is little understanding of how teams of cyber defenders operate, and what team processes and communications lead to more effective team performance. Likewise, research is needed that addresses the composition of teams and particularly, provides insight into what kinds of people are needed and how to best cope with situations where highly talented individuals are disinclined and lack the skills needed to operate in a team context.

3 R&D Addressing the Human Dimension in Cyber Operations

Workshop participants were divided into four groups who developed somewhat overlapping research proposals. The products of the four groups have been integrated to emphasize those points where there was a common appraisal of the problem and the corresponding research questions.

3.1 What Is the Problem and Why Is It Hard?

Today, the cyber defender is placed in an untenable position. They are asymmetrically disadvantaged faced off against a continually evolving opponent who can attack anywhere, anytime. The boundaries of the battlespace are ill-defined, both temporally and spatially.

Ground truth regarding the attacker, what they've done and how they've done it is rarely known with certainty. Any solution must function within the context of an overall system that includes a broad range of users and may span organizational boundaries. In the absence of ground truth, there are no real measures of success or progress rendering the domain an art, precluding the science that might otherwise provide a basis for engineering systems solutions.

3.2 What Are the Limits of Current Practice?

Today, extensive investments are being made ad hoc to develop software tools that are intended to help cyber defenders. Actions being taken are largely short-term and reactive to known threats. There exists a relatively small pool of qualified professionals with the assignment of personnel to cyber positions often driven more by expediency than thoughtful selection.

Current measures provide little insight into the human dimension making it difficult to assess performance, much less draw conclusions regarding what is and what is not working, or the differential contribution of various factors to individual, team or organizational success. Using the tools available to them today, cyber defenders must process large volumes of high-tempo data with it uncertain that this is the right data or that the data is being used in the right way, given that we do not have a good understanding of the actual work being done. Finally, there has been an insufficient allocation of resources to enable long-term strategic solutions that may require structural and organizational change.

3.3 What Are the Objectives and What Difference Will It Make?

A coordinated R&D program is needed to accomplish three separate objectives.

The first objective is to conduct human factors analysis and scientific studies to establish foundational knowledge concerning factors underlying the performance of cyber defenders. These studies should address a range of pertinent issues that include:

- The roles of defenders, users, adversaries, policy makers and the public, providing an extensible collection of use cases;
- The different jobs and functions within cyber defender teams and the associated knowledge, skills and abilities needed to fulfill these functions;
- Cognitive processes involved in typical tasks and associated measures of performance both as a basis for selection, and training and operational performance assessment;
- Methods and materials for training to both requisite levels of performance, as well as a progression from proficient to expert, and potentially elite performer.
- Allocation of functions between humans and machines, including opportunities to augment human performance through specific technological developments.

The second objective involves the development of models that capture key processes that mediate interactions between defenders, users, adversaries and the public. Models should provide sufficient complexity to enable experimentation concerning alternative tactics, techniques and policies. Models should also accommodate insertion of alternative technologies, enabling estimates of the relative returns on investment.

The third objective is to develop a multi-purpose test environment for conducting controlled experiments that enables systems and human performance measurement. The test environment should be flexible to accommodate a range of threats, software tools, modes of training, and policies, as well as mechanisms to simulate users, including the public.

Through accomplishing these objectives, cyber operations may be transformed from an art to a science, and based on that science, systems solutions may be engineered to address a range of situations. Likewise, there is an opportunity to move beyond the current state where key decisions (e.g. personnel assignment) are made on a largely ad hoc basis to a state in which there exist institutionalized processes for assuring the right people are doing the right jobs in the right way.

These developments lay the groundwork for emergence of a professional class of cyber defenders with defined roles and career progressions, with higher levels of personnel commitment and retention. Finally, operationally, the impact should be evident in improved performance, but also a transition to a more proactive response in which defenders have the capacity to exert some measure of control over the battlespace.

3.4 What Are the Measures of Success/Progress?

The first measure of success will be an ability, which does not exist today, to actually measure success. Given the primary product will be knowledge, a second measure of success will be the adoption and institutionalization of the resulting knowledge in establishing selection criteria, measures of performance, training requirements, system specifications for technology products and other related applications. A third measure of success will be the utility attributed to models and resources for conducting testing as evidenced by the amount and diversity of their use.

4 Conclusion

This paper outlines the need for R&D to address the human dimension in cyber operations. The objective of the workshop was to collect a broad set of perspectives and synthesize those perspectives in a form that may be used by different organizations to develop R&D programs.

Based upon this exercise, organizations may craft their own proposals having the benefit of knowing how other organizations view the problem and imagine the solutions. It is the intent that this broader awareness will facilitate a more coordinated effort across government organizations than would occur otherwise.

There is a rich collection of experiences in which different domains have taken concrete measures to address the human dimension within their operations. These experiences encompass both engineering analysis, scientific study and the development of technologies, practices, design guidelines and other related products.

Cyber is a relatively new domain and recognition of the human dimension in cyber operations is only now rising to the forefront. While cyber does not enjoy the wealth of knowledge and experience that is present with other domains, there is the opportunity for cyber to leverage the knowledge and experiences of these other domains to take similarly effective measures.

Acknowledgement. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Integration of Psycognitive States to Broaden Augmented Cognition Frameworks

Karmen Guevara

Computer Laboratory, Cambridge University, United Kingdom
karmen.guevara@cl.cam.ac.uk

Abstract. Augmented Cognition technologies focus on assessing and monitoring the user to produce a composite picture of their cognitive state. This is based on the mental processes of the user involving perception, memory, judgment and reasoning. It does not include the emotional, volitional or the subconscious processes. Due to the absence of input from a core human dimension – the subconscious - it is inevitable that the picture to emerge from this data will be incomplete. The focus of this paper therefore, is on this subconscious dimension. The objective is to illustrate how subconscious processes can shape behaviours and determine individuals' strategic actions. We argue that in order to formulate a complete portrait of an individual's cognitive state, it is important to integrate the subconscious dimension.

Keywords: Psycognition, characterology, subconscious behaviours, character strategies, critical incident breakdown, situational appropriate behaviour, inner subjective domain.

1 Introduction

Augmented Cognition technologies focus on assessing and monitoring the user to produce a composite picture of their cognitive state. This is based on the mental processes of the user involving perception, memory, judgment and reasoning. It does not include the emotional, volitional or the subconscious processes. Due to the absence of input from a core human dimension – the subconscious - it is inevitable that the picture to emerge from this data will be incomplete. The focus of this paper therefore, is on this subconscious dimension. The objective is to illustrate how subconscious processes can shape behaviours and determine individuals' strategic actions. We argue that in order to formulate a complete portrait of an individual's cognitive state it is important to integrate the subconscious dimension.

2 Focus and Objectives

Thus the key focus of this paper is the subconscious dimension. Our objective is to illustrate how subconscious processes shape behaviours and determine the strategic

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 423–432, 2013.

actions of individuals. The aim is to demonstrate how the inner subjective dimension can contribute to the understanding of individual cognitive states and therefore, lead to augmented cognition systems that can expand and enhance this state. An integrated view of the cognitive state and the interrelationship between the key four components is diagrammatically outlined in Figure 1.

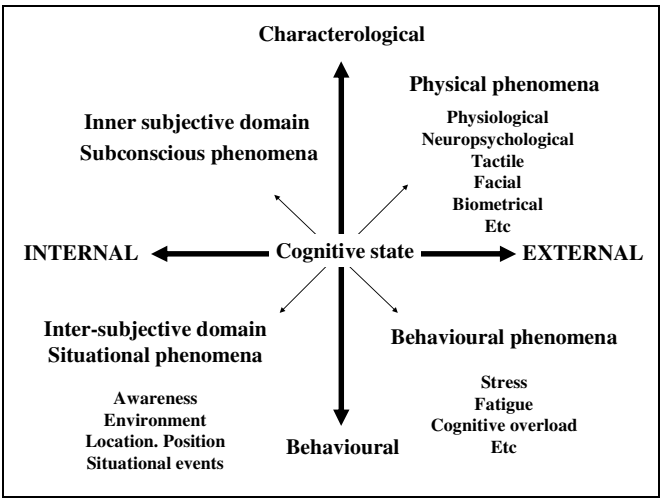


Fig. 1. An Integrated Perspective

Subconscious phenomena are embedded in the inner subjective domain. Since these are usually triggered by external events, the interrelationship with situational phenomena is important. The interaction between the two produces the manifestation of visible or measurable physical and behavioural phenomena. The inner subjective domain, consisting of subconscious phenomena is pivotal to the formation of composite pictures of current cognitive states. It is the subconscious phenomena that create the landscape of the inner subjective domain and determines how individuals experience the external world, the unconscious habitual, behavioural patterns and the decisions and strategies adopted in response to external stimuli.

Thus our hypothesis is by delving into this subconscious domain, it becomes possible to develop richer and more composite views of individual cognitive states.

3 Theoretical Framework

We refer to the inner subjective domain as the *Psycognitive domain,* which will be examined within a theoretical framework drawn from Psycognition. Psycognition is based on the theory that external behaviours arise from the subconscious that shapes how individuals perceive and experience the world. The research methodology derived from Psycognition is focused on understanding the subconscious processes associated with the semi-predictable emotional responses and behaviours that reside in the subconscious.[1]

The concept of characterology lies at the core of the Psycognition theoretical framework. Characterology refers to the set of core beliefs formulated early in an individual's development and the behaviours that are predicated on these beliefs.[2] Psycognition applies a framework of characterological types to examine subconscious processes and behavioural strategies. An overview of these is provided in Table 1. For purposes of clarity, colloquial terms are used instead of clinical ones. Despite the variations in terminology the framing of character types is similar to other topologies.[3, 4]

The identification of a dominant character orientation provides a basis for developing hypotheses about the interdependency between subconscious core material and behaviourial strategies. From this we can begin to formulate approximate predictions of how individuals will respond in certain situations. This is the framework for our examination of the interrelationship between the Psycognitive inner subjective domain and external behaviours in the following sections.

Table 1. An Overview of Characterological Themes

Character Position	Behavioural Orientation	Core Belief
Mr Safety	Safety - trust	The world is dangerous
Mr Action	Performance - recognition	Self worth = achievement
Mr Endurance	Indirect control-endurance	Not good enough but do the best
Mr Freedom	Freedom - Be the best & win	In charge – power - control
Mr Self-Reliant	Challenge. Going it alone	Never rely on others. Self-care
Mr Expressive	Attention, avoid separation	Not interesting-not listened to

4 Examination of the Psycognitive Inner Subjective Domain

Our examination of the interrelationship between the Psycognitive inner subjective domain and external behaviours is drawn from a research study of the strategies and decision making of a sample of RAF fighter pilots.[1] The investigation focused on how subconscious processes influenced the pilots' behaviours in handling critical incidents. The line of enquiry was whether there were significantly different characterological orientations among the research sample and if so, how these differences influenced the subjects handling of critical incidents.

Examples from the research illustrate how characterological orientation influences the strategies employed in three kinds of critical incidents:

1. *Information Overload:* The eight adjustment processes to information overload have been typified as: omissions, errors, filtering, abstracting, multiple channels, queuing, escape and chunking.[5]
2. *Control Breakdown:* a perceived or actual breakdown of control in a situation.
3. *Plan Breakdown:* a plan cannot be carried out, therefore objectives cannot be reached.

In the following sections we draw from the sample of fighter pilots to examine the differences in individual characterologies in relation to the strategies the subjects adopted in each critical incident.

5 Strategic Behaviours – Information Overload

The research findings pointed to differences in the subjects' adjustment processes in information overload incidents. The four strategies consistently applied in the data are presented in Table 2.

Table 2. Information Overload Strategies

Strategy	Adjustment Process	Cognitive & Behavioural Response
A	Filter, chunk, escape	Goes for a lot of information in an attempt to control overload Quantity is important – he determines the quality Core belief around not trusting dominates behaviour Withdrawal from the situation
B	Increase speed	Speeds up & goes faster Goes for a lot of detail Core belief around, "what more do I need to do here?"
C	Queue & delay	Slows things down Timing is important to receive, consider & respond Core belief around doing his best & waiting for the outcome
D	Abstraction, manipulation of multiple channels	Attempts to deflect the situation through abstraction Manipulates situations to maintain control Core belief around, "I can handle this."

Each strategy is quite different, for example, the adjustment process of Strategy A is based on controlling the overload by gathering as much information as possible, chunking it into pieces of manageable size and filtering it for quality. In contrast, Strategy D deflects the situation through abstraction and by drawing upon multiple channels of information, instead of relying only on one. The manipulation of the information is a form of maintaining control. Further contrasts are evident in Strategies B and C. Strategy C adopts the approach of slowing things down, queuing and delaying information. While Strategy B is based on speed – i.e. of obtaining as much information as possible in order to obtain more detail.

6 Strategic Behaviours – Control Breakdown

The findings highlighted pronounced differences in the subjects' strategic behaviours in control breakdown situations. These are outlined in Figure 2. Control Strategy B is orientated around values and performance where control is relinquished only after a considerable effort to understand the breakdown and when the subject is certain that

his values are not being compromised. This contrasts with the focus of Strategy D on power and maintaining control, where the subject superimposes whatever he can to maintain control over the situation.

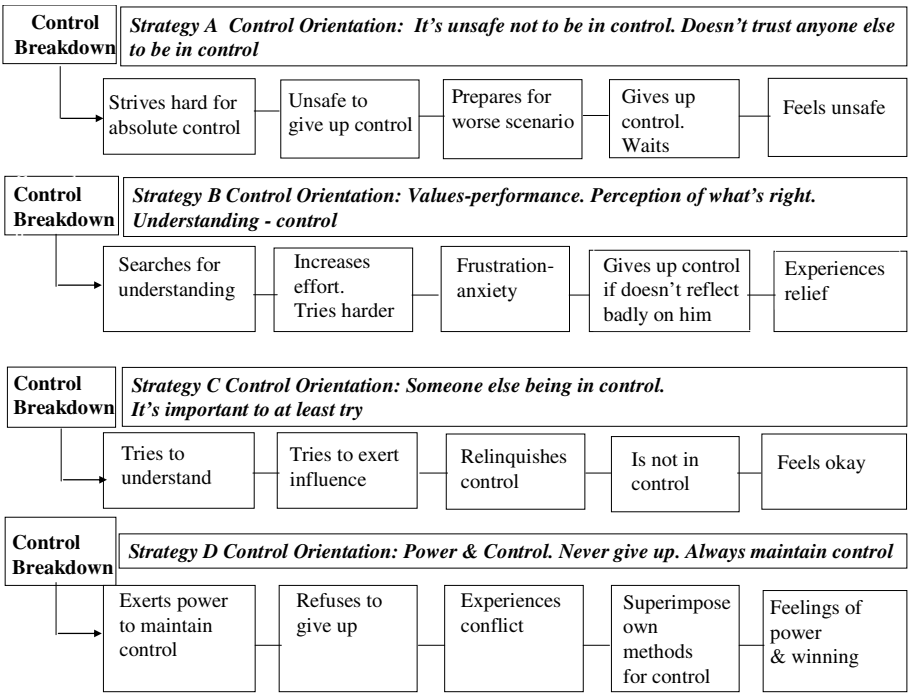


Fig. 2. Strategic Control Orientations

There were also differences in how the subjects' experienced the loss or recovery of a control breakdown. The subject with control Strategy B initially experiences frustration and anxiety, but is relieved when he finally has to relinquish control. This differs from the subject with control Strategy A, who experiences a lack of safety and therefore prepares for the worse. This contrasts with the control Strategy D subject who experiences feelings of power and winning when he refuses to relinquish control.

7 Strategic Behaviours – Plan Breakdown

Further evidence of differences in strategic behaviours was also found in the subjects' handling of breakdown incidents. The factors provoking a breakdown varied among the subjects. For example, the breakdown factor in Strategy A is a violation of values, for Strategy B it is the compromise of values and principles and for Strategy C it is the failed plan. There is no breakdown factor for Strategy D since the tactical plan is changed to enable the achievement of the goal. A divergence also emerged in the

subjects' final response to not achieving their plan. There was a range of responses from a complete withdrawal, a refusal to accept, creating a rationale for acceptance, through to compromise. These strategic differences are shown in Table 3.

Table 3. Plan Breakdown Strategies

Strategy	Primary Motivator	Goal	Strategic Tactics	Breakdown Factor	Response
A	Performance	Achievement	Goal focus Manipulates Persists	Violation of core values	Withdraws
B	Values Principles	Achieving what's right	Focus on what is right Perseverance Tactical change	Compromise on values & principles	Battles system Accepts failure if rationalized
C	Be the best possible	Influence the outcome	Provides input to situation	Failed strategy Waits	Compromise
D	Success & Winning	Achieving the goal	His goal - his way. Impulse over-rides rational thinking & judgment	None – success driven. Achievement only	Changes plan Refuses to accept failure

8 Themes and Patterns

The findings indicated that the subjects tended to apply the same strategy to the different incidents. For example, the focus on trust and safety in Strategy A was applied to both the information overload and control breakdown incidents. This was also the case for Strategy B where the focus on speed and effort applied to both of these incidents, as well as for Strategy C with the focus on slow and delay. The theme for Strategy D – control and manipulate was evident in all three incidents. These parallels suggest that the subjects organise their experiences and behavioural responses around certain underlying core beliefs. For Strategy A it is safety and trust; for Strategy B, values and performance; for Strategy C – to do one's best and compromise and for Strategy D, control and manipulation. These patterns are shown in Table 4.

Table 4. Themes and Patterns in Strategic Behaviours

	Information Overload	Control Breakdown	Plan Breakdown
A	Doesn't trust information Controls with quantity	Unsafe not to be in control Unsafe if gives up control	Manipulates situation Withdraws if values are compromised
B	Speeds up Focuses on detail	Increases effort – tries harder Gives up if not a negative reflection	Focuses on what is right Accepts it if it's rational
C	Slows and delays things Does his best and waits	Attempts to understand Relinquishes control	Focus on doing his best Adapts and compromises
D	Deflects through abstraction	Exerts power to control Imposes own methods	Impulse overrides rational thinking –refuses to accept failure

9 Strategic Behaviours in Terms of Characterology

An examination of the subjects' behaviours in terms of characterology drew close parallels between the characterological orientations outlined in Table 1 and the subjects' strategies. The theme of safety and trust for Strategy A corresponds to 'Mr Safety' and the performance and action focus of Strategy B corresponds to 'Mr Action'. There is also a parallel between 'Mr Endurance' and the orientation of Strategy C to do one's best and wait. The theme of power, control and freedom for Strategy D closely relates to 'Mr Freedom'. A mapping between the subjects' strategies and characterological orientation is provided in Table 5. This has been extended to include the strategic behaviours and the barriers associated with each characterology.

Table 5. Mapping of Subjects' Strategies with Characterological Orientation

Strategy A: Safety - Trust	Mr Safety	Overly focused on detail & analysis Excludes external factors Creates own world	Information overload: Escape, chunk, filter	**Barrier: Insight** Withdrawal Confusion leads to fear & withdrawal
Strategy B: Performance Action	Mr Action	Orientated around action & performance Focus on the logical & rational	Information overload: Speed up	**Barrier: Response** Action Lack of satisfaction leads to frustration & more action
Strategy C: Influence - Endurance	Mr Endurance	Bearing up; delaying; resisting others & waiting	Information overload: Queuing & delaying	**Barrier: Response** Resistance Feels pushed – resists
Strategy D: Power - Freedom	Mr Freedom	Ensures own choices & decisions Seeks adventure & excitement Being in charge	Information overload: Abstraction, multiple channels, manipulation	**Barrier: Insight** Deception Invests in image – leads to deception

The findings suggest the subjects' characterology emerges at points of breakdown when subconscious behaviours arise and begin to direct the strategic behaviours. For this reason, we see one subject handling an information overload situation by seeking large quantities of information through a lack of trust, another speeding up incoming information and yet another slowing down and delaying information.

Characterology also provides an explanation for the differences that appeared in the subjects' handling of control breakdown situations. For example, there was a significant variation in how the subjects responded to a breakdown in control, ranging

from at one end of the spectrum feeling unsafe when control is lacking, then to giving up control only if values and performance are not compromised, and then relinquishing control only if it can be rationalised and finally to the opposite end of the spectrum - a refusal to relinquish control.

The analysis of the subjects' behaviour in the data identified four of the six characterological types outlined in Table 1. This was fortuitous. Each of the four subjects' strategic behaviours in the three critical incidents corresponded to a different characterological type - 'Mr Safety, Mr Action, Mr Endurance, Mr Freedom'. The limited size of the sample did not provide the scope for the two remaining characterologies to emerge.

10 Barriers and Situational Appropriate Behaviours

Barriers are habitual behaviours that block the process of taking appropriate actions and decisions**.** Barriers essentially are defences against failure, which reverse the appropriate behaviour that needs to occur.[2] Each characterology has a disposition towards a particular barrier as shown in Table 5. For example, Mr Safety tends to have an insight barrier leading to a withdrawal response. Mr Action is disposed towards a response barrier.

The cycle for situational appropriate behaviours consists of four main functions: clarity, response, effectiveness and insight. Our research highlighted the interruption of two of these functions in breakdown situations.[1,6,7] The evidence indicated a number of incidents where the clarity function was interrupted by an insight barrier and the effectiveness function by a response barrier. Both of these functions are essential to situational appropriate behaviours.

The process begins with the clarity function, which arises from awareness, attention and information. With an absence of insight there isn't the clarity with which to move forward to the next function. As illustrated in the data, when this occurs some strategies involved a process of continually seeking clarity by gathering more information or through the elicitation of input.

There were other examples in the data of the interruption of the effectiveness function by a response barrier. When this barrier emerged in some of the strategies, the subjects experienced difficulty in responding with appropriate and effective action and responded instead by withdrawal, resistance or rebellion.

11 Characterology as a Dominant Behavioural Force

A new research question emerged when the data pointed to evidence of counteractions of normal characterological tendencies by the military system. Under normal circumstances the military system - the culture, ethos, rules and regulations, provides for automatic behaviour. When this system is the dominant force, individual thinking and behaviours spontaneously draw upon this system. When the strength of the system is in the forefront, the core material underlying character strategy resides in the background.

The data was examined to determine whether there were incidents where the subjects' characterology became the dominant force in driving their behaviour and strategies and the military system receded into the background. A number of specific incidents were highlighted in the data where the strength of the subjects' characterology superseded the military system. For reasons of confidentiality these specific incidents cannot be described. However, the generic behaviours that led to the superimposition of characterological behaviours over the military system are illustrated in Table 6.

The strategic behaviours evident in the data suggest that under certain circumstances such as, extreme high stress and life threatening situations or involving a compromise of character in terms of values and principles, the strength of an individual's characterology will emerge and counteract the military system. In the majority of cases however, the evidence pointed to the predominance of the military system over the subjects' normal character tendencies. This is an indication that the subjects were psychologically well balanced and fully integrated into the military system.

Table 6. The Dominance of Characterology

Strategy	Cognitive Interference/ Character Compromise	Military 'System' in Background
A	Safety & trust	Will not rely on 'the system' for safety Has his own rules for safety
B	Performance & recognition	Bypasses the system if necessary to maintain character integrity Will not allow the system to undermine his performance
C	To do one's best & influence	Bypasses the system if necessary
D	Power, control, freedom, adventure	Dangerous situations; overrides rules A disregard for procedures

12 Research Conclusions

An important finding from this research is that despite the similarities in the subjects' background, training, experience and the strength of the military culture, there were significant differences in how they responded to critical incidents involving a breakdown. Our conclusion here is that the differences in the strategic behaviours can be attributed to how the subjects organise their experiences around the subconscious material rooted in their characterological orientation. This is more pronounced in situations where the subjects are faced with an actual or perceived threat to survival, experience a breakdown in plan or control, an overload of information or if values and principles are compromised.

These conclusions challenge the common assumption that individuals draw upon conscious and rational behaviour. The findings indicate this is not valid in all circumstances. A key conclusion is that this research provides evidence that the Psycognition characterological framework has diagnostic and potential predictive power.

This can provide us with a basis for developing an understanding of the subconscious strategic behaviours that could emerge in critical incidents and in other challenging environments.

13 Implications

Understanding behaviour under severe stress has key implications for augmented cognition technologies and system design in general. Military systems are just one of the many kinds which can serve to provide high risk operational conditions which demand that human stress behaviours be studied. The results of this research suggest that under extremes of stress, subjects organise their experiences and behavioural responses around certain underlying core beliefs. Under certain conditions the subjects will depend on these deeply embedded core beliefs in the subconscious to guide their decisions and action, *as well as, or more than, rational reasoning*. Moreover, when drawing on their experience to compensate, the subjects organise those experiences around the subconscious material rooted in their characterological orientation.

Thus systems design based on the assumption that operational staff draw only upon conscious and rational behaviour may be flawed. From the evidence given here, augmented cognition technologies need to include a *Psycognition characterological framework*, to provide a key diagnostic dimension for understanding strategic behaviours driven by the subconscious that are manifested under stress, i.e. in crucial moments of decision taking in life-threatening or other conditions where individuals are under intense strain.

References

1. Guevara, K.: Psycognition: An Exploration of the Strategic Behaviours Underlying Fighter Pilots' Decision Making in Critical Incidents. Technical Report, DERA, CHS, Farnborough, UK (1997)
2. Kurtz, R.: Body-Centered Psychotherapy. Life Rhythm, California (1990)
3. Sharp, D.: Jung's Model of Typology. Inner City Books, Toronto (1987)
4. Shapario, D.: Neurotic Styles. Basic Books, New York (1965)
5. Miller, J.G.: Living Systems. University Press of Colorado, Colorado (1995)
6. Guevara, K.: Cognitive Architectures for Supporting Strategic Behaviours in Adaptive Systems. In: 4th Joint GAF/RAF/USAF Workshop on Human-Computer Teamwork, HE Crew Conference Report, pp. 91-95 published by AFRL Human Effectiveness Wright Patterson AFB: AFRL-HE-WP TR 1999-0235 (1997)
7. Guevara, K.: Psycognition: Cognitive Architectures for Augmented Cognition Systems. In: Stephanidis, C. (ed.) Posters, Part I, HCII 2011. CCIS, vol. 173, pp. 275–279. Springer, Heidelberg (2011)

Human Performance Assessment Study in Aviation Using Functional Near Infrared Spectroscopy

Joshua Harrison[1], Kurtulus Izzetoglu[1], Hasan Ayaz[1], Ben Willems[2], Sehchang Hah[2], Hyun Woo[2], Patricia A. Shewokis[1,3], Scott C. Bunce[4], and Banu Onaral[1]

[1] School of Biomedical Engineering, Science & Health Systems, Drexel University,
[2] Atlantic City International Airport: Federal Aviation Administration
W.J. Hughes Technical Center
[3] Nutrition Sciences Department, College of Nursing and Health Professions, Drexel University
[4] Penn State Hershey Medical Center and Penn State College of Medicine
{jlh444,ki25}@drexel.edu

Abstract. Functional near infrared (fNIR) spectroscopy is a field-deployable optical neuroimaging technology that provides a measure of the prefrontal cortex's cerebral hemodynamics in response to the completion of sensory, motor, or cognitive tasks. Technologies such as fNIR could provide additional performance metrics directly from brain-based measures to assess safety and performance of operators in high-risk fields. This paper reports a case study utilizing a continuous wave fNIR technology deployed in a real-time air traffic control (ATC) setting to evaluate the cognitive workload of certified professional controllers (CPCs) during the deployment of one of the Federal Aviation Administration's (FAA's) Next Generation (NextGen) technologies.

Keywords: Near-infrared spectroscopy, optical brain imaging, fNIR, human performance assessment, air traffic control, workload.

1 Introduction

Military and civilian aviation personnel are increasingly required to utilize larger and more complex automation systems. Hence, the information-processing load and decision-making demands have recently been increased on aviation personnel including pilots and air traffic controllers. While skilled operators have demonstrated the ability to sustain a sufficient level of performance as task difficulty increases, eventually increased workload leads to a decrease in performance that ultimately can lead the controller to make very dangerous or even deadly errors [1]. As new technology is implemented to increase the safety of air travel it is imperative to avoid adversely affecting the controller's performance by overloading the controller with the technology. Emerging wearable functional brain activity monitoring technologies can help evaluate the cognitive status and capacities of the crew in cockpit as well as in ground control stations. Such technologies could become an important asset in maintaining safe and effective performance through providing additional performance metrics

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 433–442, 2013.

directly driven from brain-based measures. Functional near infrared (fNIR) spectroscopy is a field-deployable non-invasive optical brain imaging technology that measures cerebral hemodynamics within the prefrontal cortex in response to sensory, motor, or cognitive tasks [2-4]. This paper aims to introduce research efforts underway to progress fNIR technology towards field applications in aviation including a study with the Federal Aviation Administration (FAA). In collaboration with the FAA's William J. Hughes Technical Center, we explored the impact of alternative Conflict Resolution Advisory (CRA) conditions on air traffic controller (ATC) behavior and workload.

1.1 Principles of fNIR in Cognitive Workload Assessment

Changes in cognitive workload are known to cause a predictable response in neurophysiological and psychophysiological variables [5]. As neurons are differentially activated according to task, there is a subsequent change in cerebral blood flow to match the metabolic demand of the neurons, a phenomenon known as neurovascular coupling. Similarly, as the demand of a neuron is increased there is a local increase in oxygenated hemoglobin (HbO_2) and a decrease in deoxygenated hemoglobin (HbR) indicating increased brain metabolism. During increased brain activity, the required local oxygen supply is generally overestimated by the neuron resulting in an increased level of cerebral blood oxygenation [6]. Since HbO_2 and HbR have distinctive optical properties in the near-infrared light range, the relative change in concentration of these molecules during differing levels of brain activation can be measured using optical methods (Fig. 1) [7].

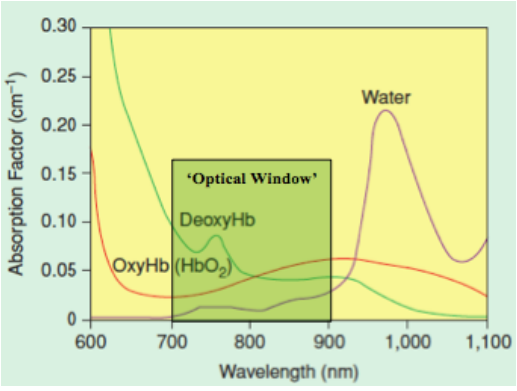


Fig. 1. Absorption spectra of HbO_2, HbR and H_2O at different wavelengths of light [8]

Chance et al. [9] first reported oxygenation changes related to brain activity in the prefrontal cortex using a fNIR spectroscopy based system during a problem solving task . Prior to utilizing objective brain based measures for cognitive workload assessment, subjective measures of workload were required that often included secondary tests that hindered the operator's ability to perform the task. Our laboratory recently demonstrated the ability of fNIR to assess brain activity and its relation to

human performance assessment such as the cognitive workload monitoring of ATCs, task complexity, skill acquisition, problem solving, learning/training assessment, and ATCs controlling fixed numbers of aircraft and military personnel commanding fixed numbers of warships [2, 10-16].

1.2 Physical Principles of Optical Brain Imaging

When near infrared light, with a wavelength of 700-900 nm, enters brain tissue, much of the light is scattered, some is absorbed and a small portion is reflected back to the sensor [17]. Water, and thus tissue, does not absorb light very highly in this range, however, hemoglobin (HbR) and Oxy-hemoglobin (HbO_2), have distinct spectra within this 'optical window', which makes it possible to detect changes in HbO_2 and HbR concentration through spectroscopic techniques (Fig. 1) [8, 18]. Utilizing the peak absorption wavelengths of HbR and HbO_2's chromophores, 730 nm and 850 nm respectively, it is possible to measure the relative changes of both HbO_2 and HbR to effectively monitor the brain activity of individuals with fNIR spectroscopy brain imaging [2, 8, 17]. Applying the modified Beer Lambert Law, based off of these principles, the relative changes in HbO_2 and HbR concentrations compared to a baseline measurement can be calculated.

$$OD_\lambda = \log\left(\frac{I_{in}}{I_{out}}\right) \approx \varepsilon_\lambda \cdot c \cdot d \cdot DPF_\lambda + G \tag{1}$$

The parameters for this equation are as follows: OD_λ is the optical density at a specific wavelength, I_{in} is the intensity of the inputted light, I_{out} is the intensity of the detected light, ε_λ is the extinction coefficient of the two chromophores, either HbO_2 or HbR, d is the distance the light traveled, ~1.25 cm with our device, DPF is the differential path length factor due to high scattering, and G is the attenuation factor. This modified law depends on the theory that near infrared light traveling through the tissue is scattered at a constant level, allowing "G" and "DPF" to be considered constants in the equation [19].

Additionally, when the light intensity is kept constant and measurements are taken over two different time periods and two different wavelengths the equation reduces to:

$$\begin{bmatrix}\Delta OD_{\lambda 1}\\ \Delta OD_{\lambda 2}\end{bmatrix} = \begin{bmatrix}\varepsilon_{\lambda 1}^{HbR} d \cdot DPF_{\lambda 1} & \varepsilon_{\lambda 1}^{HbO_2} d \cdot DPF_{\lambda 1}\\ \varepsilon_{\lambda 1}^{HbR} d \cdot DPF_{\lambda 2} & \varepsilon_{\lambda 2}^{HbO_2} d \cdot DPF_{\lambda 2}\end{bmatrix}\begin{bmatrix}\Delta c^{HbR}\\ \Delta c^{HbO_2}\end{bmatrix} \tag{2}$$

Exploiting equation 2, the relative change in concentration of HbR and HbO_2 (ΔC_{HB} and ΔC_{HBO2}) can be deduced. Subsequently, oxygenation (Oxy) and total blood flow (HbT) can be calculated from ΔC_{HB} and ΔC_{HbO2}:

$$\begin{aligned} Oxy &= \Delta C_{HbO_2} - \Delta C_{HbR}\\ HbT &= \Delta C_{HbO_2} + \Delta C_{HbR}\end{aligned} \tag{3}$$

1.3 Continuous Wave fNIR Device

The continuous wave fNIR device deployed for this study was first implemented by Chance et al. [20], further developed at Drexel University, and manufactured by fNIR Devices LLC (Potomac, MD, www.fnirdevices.com). The device consists of a sensor pad with 4 light emitting diode (LED) light sources and 10 detectors, a portable hardware box, and a laptop computer (Fig. 2).

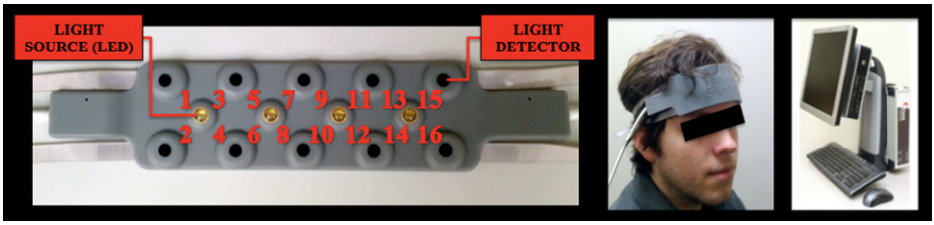


Fig. 2. Design and positioning of portable fNIR device. 16 channel portable fNIR sensor pad design with all 16 channels labeled (Left). Positioning of the fNIR sensor pad on a subject's head (Center). fNIR data acquisition box and Computer system (Right).

1.4 FAA's Next Generation (NextGen) Implementation

The Next Generation Air Transportation System (NextGen) includes an array of new technologies and ideas to prepare the National Airspace System (NAS) for increased air traffic while, at a minimum, maintaining current levels of safety and efficiency. Providing controllers with tools that aid in the determination of the optimal solutions will make the system more predictable and ultimately increase the efficiency of the system. One possible tool tested within this experiment is the CRA, which provides ATCs with strategic conflict detection and trial planning capabilities (Fig. 3).

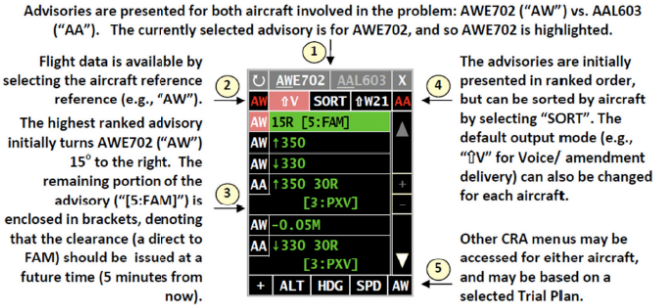


Fig. 3. Potential layout of a search all CRA Menu

Given the complexity of the CRA and the concern that controllers may change the way in which they perform their job, consideration has been given to the implementation of the system. ATCs work in pairs while directing traffic. Within the pair, one controller is referred to as the radar controller (R-side) controller and the other

controller is the data controller (D-side), a position often referred to as the radar assistant. For the implementation of the CRA, three conditions were explored: (1) neither controller had access to the CRA (baseline); (2) D-side only had access to the CRA (D-Only); (3) both controllers had access to the CRA (Both).

2 Method: Brain Activity Monitoring During ATC Simulation

The airspace used in the experiment consisted of two active high altitude sectors, 20 and 22, of Kansas Center (ZKC; Fig.4). Traffic scenarios were developed based on samples extracted from the Aircraft Situation Display to Industry (ASDI) feed to ZKC. The traffic was filtered to include only aircraft that crossed a volume of airspace of 300 by 300 nautical miles that included the sectors used in the experiment. Four simulation pilots were utilized for each controlled sector to simulate real controller pilot communication.

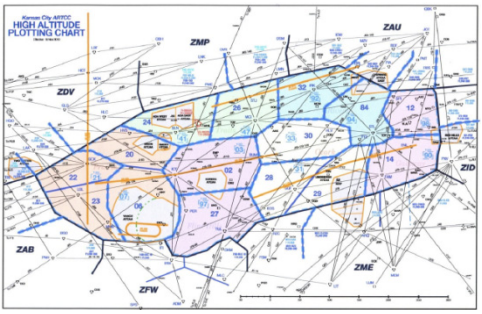


Fig. 4. High altitude sectors of ZKC

The simulation was conducted at the Research Development and Human Factors Laboratory (RDHFL), William J. Hughes Technical Center, Atlantic City International Airport, NJ. In the experiment room, there were two pairs of controller workstations, with fNIR data collected from one pair of controllers (Fig. 5). The prefrontal cortex of all participants was monitored during the air traffic control simulations with a portable, 16-channel continuous wave fNIR system (Fig. 2). The fNIR device was placed underneath an occulometer device for data collection. The fNIR recording was synchronized with the traffic scenarios through implementation of a custom application implemented to send event markers to the fNIR data acquisition computer via RS232.

The Distributed Environment for Simulation, Rapid Engineering, and Experimentation (DESIREE) was utilized to simulate the En Route Automation Modernization (ERAM) system. The DESIREE emulation of the ERAM system was modified to accommodate the CRA. A voice communication system was utilized that mimics the operational Voice Switching and Communications System (VSCS) utilized by ATCs. This system allows for air/ground communication between controllers and simulation

pilots as well as ground/ground communications link between controller participants for inter-sector communications. The ATCs had access to two types of communication systems voice (VoiceComm) and data (DataComm) with a constant 30% of aircraft equipped with DataComm to simulate realistic near-future communication conditions.

Fig. 5. ATC simulation center: Each workstation consisted of a high-resolution (2048 x 2048), radarscope, keyboard, trackball and direct access keypad

2.1 Experimental Procedure

Prior to the study, all participants signed informed consent statements approved by the Federal Aviation Administration's Human Subjects Review Board. Twelve certified professional controllers (CPCs), previously unfamiliar with ZKC airspace, volunteered for the study. Prior to beginning the study, participants received 1 day of extensive training on the airspace, systems and procedure. During this day, the volunteers participated in 5, 30-minute training scenarios. The first training scenario used a low traffic level (33% to 66% of the monitor alert parameter (MAP)). A higher traffic level (33% to 100% of the MAP value) was utilized for the second training scenario. The final training scenario used a traffic volume that was high as the traffic volume of the experimental scenario (33% to 150% of the MAP value). MAP is previously described as the number of aircraft that a sector/airport can accommodate without degraded efficiency during a specific period of time [21].

Over a three-day test period, the participants completed a total of 9 test sessions. Each day, 3 practice and 3 test sessions were completed utilizing each of the three conflict resolution advisory implementations, baseline, D-side, and both. Each training session was performed for 30 minutes, with the previously described traffic levels, and a 30-minute break given prior to test scenarios. The training scenarios were implemented to allow the controllers to familiarize themselves with the CRA version that would be available during the test scenario. Each test scenario lasted for 50 minutes, in which the traffic volume was ramped from 33% to 150% of the MAP value. All training and test blocks were counterbalanced to minimize the order effects across participants.

2.2 Data Acquisition

Throughout the entire experiment physiological data and subjective workload assessment data were collected. fNIR sensor recordings were acquired with a sampling rate of 2 Hz and workload assessment keypad (WAK) ratings were made by the controllers once every 2 minutes. Eye-tracker data was also collected throughout the experiment on all subjects. A flexible fNIR sensor (Fig. 2) consisting of 4 light sources and 12 light detectors, with peak wavelengths at 730 nm and 850 nm, was placed over each subject's forehead to scan the prefrontal cortex. This source-sensor configuration generates a total of 16 prefrontal cortex measurement locations per scan (Fig. 2). COBI Studio acquisition software (©2010, Drexel University) was used for fNIR data collection. Raw light intensity recordings were low-pass filtered with a finite impulse response, linear phase filter with a cut-off frequency of 0.14 Hz and order of 20 to reduce high frequency noise. Using filtered raw fNIR measures; HbO_2, HbR, Oxy and HbT were calculated using equations 1-3. The baseline chosen corresponds to the time period when the ATCs were controlling less than 10 aircraft. A large amount of fNIR data was visually rejected due to sensor decoupling caused by subject discomfort attributed to the combination of eye tracker/fNIR.

3 Results

Each session was divided into 7 'Aircraft Count' blocks to distinguish between traffic levels under ATC command. For statistical analysis, the fNIR data from 22 sessions (Baseline, D-Only n=7; Both n=8) was analyzed with a 6 (Aircraft Levels) x 3 (CRA Conditions) repeated measures ANOVA to compare within subject factor effect of the different aircraft levels and between subject effect of CRA condition (Fig. 6).

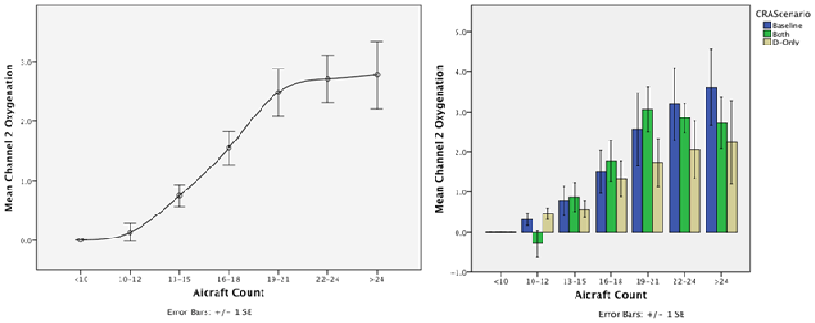


Fig. 6. Mean relative oxygenation changes with increasing aircraft count (Left). N-values for the plot are as follows (Aircraft <10, 13-15, 16-18, n = 25; Aircraft 10-12, 19-21, 22-24, n=24, Aircraft >24, n = 13). Mean relative oxygenation changes with increasing aircraft count divided by CRA Implementation (Right).

The data block '>24' was excluded from the statistical analysis due to few data points in this region. Greenhouse-Geisser was used to correct for sphericity and Tukey's post-hoc analysis was performed to identify the locus of the main effect of aircraft levels on workload. The significance criterion chosen for all statistical tests

was α= 0.05. The within subject effect of aircraft count was found to be significant (Fig. 6; $F_{5,105} = 37.441$, $p<0.001$, $\eta_p^2 = 0.663$). The between subject CRA effect was found to have no significant effect on oxygenation levels (Fig. 6). The results of post-hoc analysis indicated that the within subject effect of each aircraft count is significantly higher than the preceding blocks except for the comparison of '<10 Aircraft' and '10-12' Aircraft (Table 1).

Table 1. P-Values for Post-Hoc Analysis on Within Subject Effect of Aircraft Count. * Indicates significant differences.

Aircraft Count	<10	10-12	13-15	16-18	19-21	22-24
<10	-	1.000	0.038*	0.002*	0.000*	0.000*
10-12	1.000	-	0.000*	0.000*	0.000*	0.000*
13-15	0.038*	0.012*	-	0.003*	0.000*	0.000*
16-18	0.002*	0.000*	0.003*	-	0.002*	0.000*
19-21	0.000*	0.000*	0.000*	0.002*	-	0.001*
22-24	0.000*	0.000*	0.000*	0.000*	0.001*	-

4 Discussion

This paper presents the preliminary finding that cognitive workload of air traffic controllers can be monitored accurately for continuously and incrementally changed task difficulty levels using fNIR, a portable optical brain imaging system. Previous studies have shown that mental workload can be estimated for controlled conditions in the natural working environment of the operators [2] but had not been tested for the continuously changing task difficulties as in the current study where ATCs participated for a 50-minute session to distinguish between workload levels caused by continuously increased traffic levels. For low aircraft counts it appears that ATCs can increase their cognitive function, specifically, working memory, similar to ATCs performance on the n-back test [2]. However, as the aircraft count increased beyond 100% MAP value, the ATCs could no-longer increase their cognitive function to meet the tasks demand. While oxygenation increases were all significant, the spline fit to the data indicates that the second derivative changes for the group around the 19-21 aircraft mark indicating the controllers cannot continue to increase their cognitive function to match demand. This finding may add objective physiological validity to the MAP rating system previously described by the FAA. More importantly, the two implementations of the CRA system compared to the baseline condition did not significantly change the ATC's cognitive workload. However, the D-Only implementation of the CRA shows a possible decrease in workload at higher traffic levels even though the results were not conclusive. These findings provide some insight to possible future validations of the CRA concerning workload response. Future work, employing an increase in the sample size to reduce Type II error may help to illustrate that the CRA is not adding unnecessary workload to the ATC but instead the CRA may reduce the controller's workload when implemented correctly.

Continuously and objectively monitoring the cognitive workload of ATCs and other operators, with a portable brain-imaging device, such as fNIR, may allow for an increase in safety of air travel and other high-risk activities by ensuring the operator does not become overloaded. Additionally, an accurate objective assessment of cognitive workload may help prevent operator error and allow for appropriate intervention through predicting probable errors that can arise from work overload [22-25]. An objective workload assessment system, such as fNIR, may prove to be a valuable tool in the validation of the array of FAA's NextGen systems, such as the CRA presented in this paper.

References

1. Endsley, M.R., Rodgers, M.D.: Distribution of attention, situation awareness, and workload in a passive air traffic control task: Implications for operational errors and automation. Air Traffic Control Quarterly 6(1), 21–44 (1998)
2. Ayaz, H., et al.: Optical brain monitoring for operator training and mental workload assessment. NeuroImage 59(1), 36–47 (2012)
3. Ferrari, M., Quaresima, V.: A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. NeuroImage 63(2), 921–935 (2012)
4. Izzetoglu, M., et al.: Functional near-infrared neuroimaging. IEEE Trans. Neural Syst. Rehabil. Eng. 13(2), 153–159 (2005)
5. Fairclough, S.H., Venables, L., Tattersall, A.: The influence of task demand and learning on the psychophysiological response. Int. J. Psychophysiol. 56(2), 171–184 (2005)
6. Fox, P.T., et al.: Nonoxidative glucose consumption during focal physiologic neural activity. Science 241(4864), 462–464 (1988)
7. Jobsis, F.F.: Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science 198(4323), 1264–1267 (1977)
8. Bunce, S., et al.: Functional Near Infrared Spectroscopy: An Emerging Neuroimaging Modality. IEEE Engineering in Medicine and Biology Magazine, Special issue on Clinical Neuroengineering 25(4), 54–62 (2006)
9. Chance, B., et al.: Cognition-activated low-frequency modulation of light absorption in human brain. Proc. Natl. Acad. Sci. U S A 90(8), 3770–3774 (1993)
10. Izzetoglu, K., et al.: Functional Optical Brain Imaging Using Near-Infrared During Cognitive Tasks (2004)
11. Ayaz, H., Cakir, M.P., Izzetoglu, K., Curtin, A., Shewokis, P.A., Bunce, S.C., Onaral, B.: Monitoring expertise development during simulated UAV piloting tasks using optical brain imaging. In: IEEE Aerospace Conf., pp. 1–11 (2012)
12. Ayaz, H., Bunce, S., Shewokis, P., Izzetoglu, K., Willems, B., Onaral, B.: Using Brain Activity to Predict Task Performance and Operator Efficiency. In: Zhang, H., Hussain, A., Liu, D., Wang, Z. (eds.) BICS 2012. LNCS, vol. 7366, pp. 147–155. Springer, Heidelberg (2012)
13. Bunce, S.C., Izzetoglu, K., Ayaz, H., Shewokis, P., Izzetoglu, M., Pourrezaei, K., Onaral, B.: Implementation of fNIRS for monitoring levels of expertise and mental workload. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2011. LNCS, vol. 6780, pp. 13–22. Springer, Heidelberg (2011)

14. Ayaz, H., Shewokis, P.A., İzzetoğlu, M., Çakır, M.P., Onaral, B.: Tangram solved? Prefrontal cortex activation analysis during geometric problem solving. In: 34th Annual International IEEE EMBS Conf. IEEE, pp. 4724–4727 (2012)
15. Izzetoglu, K., Ayaz, H., Menda, J., Izzetoglu, M., Merzagora, A., Shewokis, P.A., Pourrezaei, K., Onaral, B.: Applications of Functional Near Infrared Imaging: Case Study on UAV Ground Controller. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition, HCII 2011. LNCS, vol. 6780, pp. 608–617. Springer, Heidelberg (2011)
16. Ayaz, H., Izzetoglu, K., Cakir, M.P., Curtin, A., Harrison, J., Izzetoglu, M., Shewokis, P.A., Onaral, B.: Functional Brain Activity Monitoring during Unmanned Aerial Vehicle Coordination. In: 20th IEEE Signal Processing and Communications Applications Conf., Fethiye, Mugla, Turkiye, pp. 1–4 (2012)
17. Cope, M., et al.: Methods of quantitating cerebral near infrared spectroscopy data. Adv. Exp. Med. Biol. 222, 183–189 (1988)
18. Cope, M.: The application of near infrared spectroscopy to non-invasive monitoring of cerebral oxygenation in the newborn infant, in Department of Medical Physics and Bioengineering1991, University College London
19. Obrig, H., Villringer, A.: Beyond the visible–imaging the human brain with light. J. Cereb. Blood Flow Metab. 23(1), 1–18 (2003)
20. Chance, B., et al.: A novel method for fast imaging of brain function, non-invasively, with light. Opt. Express 2(10), 411–423 (1998)
21. FAA, FAA Order 7210.3 Facility Operation and Administration, Federal Aviation Administration: Washington, DC (2007)
22. Hancock, P., Parasuraman, R.: Human factors and safety in the design of Intelligent Vehicle–Highway Systems (IVHS). Journal of Safety Research 23, 181–198 (2008)
23. Hancock, P.A., Verwey, W.B.: Fatigue, workload and adaptive driver systems. Accid Anal Prev. 29(4), 495–506 (1997)
24. Parasuraman, R., Christensen, J., Grafton, S.: Neuroergonomics: The brain in action and at work. NeuroImage 59(1), 1–3 (2012)
25. Parasuraman, R., Wilson, G.F.: Putting the brain to work: neuroergonomics past, present, and future. Hum Factors 50(3), 468–474 (2008)

Robust Classification in RSVP Keyboard

Matt Higger, Murat Akcakaya, Umut Orhan, and Deniz Erdogmus

Northeastern University,
360 Huntington Ave, Boston, MA 02115
{Higger,Akcakaya,Orhan,Erdogmus}@ece.neu.edu

Abstract. To use in the Rapid Serial Visual Presentation (RSVP) Keyboard^{TM}, a brain computer interface (BCI) typing system developed by our group, we propose a robust classification method of handling non-stationarity in the electroencephelography (EEG) data that is caused by artifacts and/or sensor failure. Considering the effect of these non-stationarities, we build a mixture data model to use as EEG evidence in the fusion with an n-gram language model to develop a robust classification algorithm. Using Monte Carlo simulations on the pre-recorded EEG data containing sections with or without intentionally generated artifacts we compare the typing performances of non-robust and robust classification methods in terms of speed and accuracy.

Keywords: BCI, ERP, Spelling.

1 Introduction

Locked-In Syndrome can isolate a person from those closest to them by taking away their ability to communicate. We focus on empowering those who are totally locked in, without control of any muscle group or eye gaze, by offering them a voice. Brain computer interfaces (BCIs) offer a promising avenue to do this. Generally, BCIs are methods which extract a person's intent through measurement of internal body signals. A common method, as we employ here, is to use the voltage of a person's scalp measured through Electroencephelography (EEG). EEG is a relatively cheap, portable, non-invasive way of measuring brain waves.

There are a number of EEG brain-phenomena which have been used to classify user intent. In motor imagery, a BCI system is designed to detect the signal generated by imaging the movement of a body part [1]. Additionally a steady state visually evoked potential (SSVEP) appears when a user is exposed to a periodic visual stimulus. Exposing a user to flickering checkerboard patterns, the induced SSVEP signals can be used to learn the user's gaze position from the frequency content of their brain waves [2].

Moreover, the EEG signals are sufficient for simple letter selection in the context of a typing algorithm for people with total-LIS. P300 signal, an event related potential (ERP) which occurs when a user is surprised by a circumstance, is commonly used for BCI spelling systems. P300 speller and Berlin BCI's Hexo Spell

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 443–449, 2013.

are well known examples of such systems [3], [4]. Different than these systems, in our approach, we utilize Rapid Serial Visual Presentation (RSVP), which presents the stimuli on the same location of the screen with temporal separation. The accuracy and speed of P300 typing systems suffer from low signal-to-noise ratio (SNR), the presence of artifacts in the signal and sensor failure and other effects that cause non-stationarity in the observed EEG signals. In this paper, we focus on a method to mitigate the influence of this non-stationarity on the typing performance.

The artifacts and/or sensor failure change the underlying distribution of the EEG data obtained from a BCI system causing a change in the optimal stimuli classification rule and degrading the system performance. Our goal is to develop a classification rule that is robust to changes in the assumed data distributions. To achieve this, we estimate the distribution of the data under different conditions, and using this distribution we develop our classification rule.

The rest of the paper is as follows. In Section 2, we explain the RSVP Keyboard^{TM}, and then in Section 3, we develop the proposed robust classification rule. In Section 4, we demonstrate our experimental results, and conclude our discussion in Section 5.

2 RSVP Keyboard^{TM}

The RSVP Keyboard^{TM}consists of four main components: visual presentation, feature extraction, language modeling and the classifier used to select a symbol.

2.1 Visual Presentation

RSVP is a presentation technique in which visual stimuli are displayed as a temporal sequence at a fixed location on the screen. An example screen snapshot from the current RSVP Keyboard prototype is given in Figure 1.In the current study, RSVP contains random permutations of the 26 letters in English alphabet, a space symbol and a backspace symbol (a total of 28 symbols to choose from). We use the term "sequence" to mean a showing of all 28 symbols. If repetition is needed, all symbols are repeated multiple times to improve classification accuracy until a preset desired confidence level or a maximum number of repetition is reached. The process of repetition of sequences to choose a single symbol is named as an epoch. In an epoch, we make the assumption that the user shows positive intent for a single symbol.

2.2 Feature Extraction

The feature extraction starts by extracting stimulus-time-locked bandpass filtered EEG signals for each stimulus in the sequence. Since physiologically, the most relevant signal components are expected to occur within the first 500ms following the stimuli, the [0,500] ms portion of the EEG following each stimulus is extracted. At this stage it is important to design bandpass filters whose group

Fig. 1. RSVP Keyboard interface

delay does not shift the physiological response to outside this interval. A linear dimension reduction is applied on the temporal signals using Principal Component Analysis in order to to remove zero variance directions (i.e. zero-power bands based on the estimated covariance). The final feature vector to be classified is obtained as a concatenation of the PCA-projected temporal signals for each channel. Regularized Discriminant Analysis (RDA) [5] is used to further project the EEG evidence into scalar-feature for use in fusion with language model evidence.

RDA is a modification of quadratic discriminant analysis (QDA). QDA yields the optimal minimum-expected-risk Bayes classifier under the assumption of multivariate Gaussian class distributions. This classifier depends on the inverses of covariance matrices for each class, which are estimated from training data. To keep the calibration phase short few training samples are acquired - especially for the positive intent class. Therefore, the sample covariance estimates may become singular or ill-conditioned for high-dimensional feature vectors, which is the case here. RDA applies shrinkage and regularization on class covariance estimates. Shrinkage forces class covariances closer towards the overall data covariance as:

$$\hat{\Sigma_C}(\lambda) = \frac{(1-\lambda)\Sigma_C + \lambda\hat{\Sigma}}{(1-\lambda)N_C + \lambda\hat{N}} \tag{1}$$

Where λ is the regularization parameter, Σ_C, N_C are the class covariance estimate and number of samples for classes $C \in \{0,1\}$ respectively. $C = 0$ is the non-p300 class. $\hat{\Sigma}, \hat{N}$ is the total covariance estimate and number of samples over all classes. Regularization is administered as:

$$\hat{\Sigma_C}(\lambda,\gamma) = (1-\gamma)\hat{\Sigma_C}(\lambda) + \frac{\gamma}{d}Tr[\hat{\Sigma_C}\lambda]I \tag{2}$$

where γ is the regularization parameter, $/texttr[.]$ is the trace function and d is the dimension of the data vector.

After regularization and shrinkage, the covariance and mean estimates for each class are used in generating a scalar feature that minimizes expected risk

under the Gaussianity assumption of class distributions. This is the log-likelihood ratio

$$\delta_{RDA}(x) = \log \frac{f_N(x; \hat{\mu_1}, \hat{\Sigma_1}(\lambda, \gamma)\hat{\pi_1}}{f_N(x; \hat{\mu_0}, \hat{\Sigma_0}(\lambda, \gamma)\hat{\pi_0}} \tag{3}$$

where $\mu_c, \hat{\pi_c}$ are estimates of class means and priors respectively; x is the data vector to be classified and $f_N(x; \mu, \Sigma)$ is the pdf of a multivariate Gaussian (normal) distribution.

2.3 Language Modeling

In letter-by-letter typing, we adopt an n-gram language models at the symbol level. These models estimate the conditional probability of a letter given by the $n-1$ previously typed letters. In this study, a 6-gram model that is trained using a one-million sentence (210M character) sample of the NY Times portion of the English Gigaword corpus. Corpus normalization and smoothing methods are described in [6]. Finally, we note that the backspace symbol is assumed to have a constant conditional probability of 0.05 and the conditional probabilities of the other symbols are normalized accordingly.

2.4 Classifier

Using the class conditional score and the language model probabilities in a naive Bayes' rule based fusion model, we compute the posterior probabilities of symbols given all the evidence. We compute these probabilities for each symbol after every sequence, and a decision is made if one symbol probability reaches a desired confidence level or number of repetitions exceeds a predefined limit.

3 Robust Classifier

In the classifier, the class conditional score distributions are used assuming that these distributions remain stationary during a typing session. However, possible changes in the distribution of the EEG data, possibly due to artifacts or sensor failure, should be incorporated in the score distribution. For example, as we also explain in Section 4, we apply our method on artifact reduction assuming artifacts as possible reasons for changes in the distribution. We introduce a variable a which describes the artifact class of a particular trial. Artifact classes include a control group (no artifacts present), eye blink, jaw movement and smiling. For use in the language model fusion, we compute the score conditional distributions for the mixed conditional score distribution as

$$P(\delta_{RDA}(x)|c) = \Sigma_i P(\delta_{RDA}(x)|c, a_i)P(a_i) \tag{4}$$

where, i is the artifact index, $c = 0$ or 1 is the class label, $P(a_i)$ is the prior for artifact a_i. For each class and artifact $P(\delta_{RDA}(x)|c, a_i)$ is computed using (2.2).

4 Experiments

Four healthy operators participated in this study. For each subject, four RSVP sessions with pre-designated targets were performed using a 16-channel g.USBamp and g.Butterfly electrodes (g.Tec, Graz, Austria) in one sitting. The second session was the control session, while the first, third and fourth sessions had the subjects produce intentional jaw movement, eye blinks, and face muscle artifacts, respectively. Subjects continued to attend to the RSVP presentation during all sessions. This data is used to build and test robust and non-robust fusion models using 10-fold cross validation as explained in Section 3

We perform Monte Carlo simulations on multiple pre-recorded calibration data sets to build kernel density estimates (KDEs) of the RDA score distribution for target symbol present and not-present conditions.

We select ten different sentences and aim to spell a phrase in each sentence (called the copy phrase task). Task difficulty is determined by requiring each letter of the target phrase to have a likelihood ratio against the highest likelihood competing non-target letter within a specified interval: (1) Hard: (0.3,0.5], (2) Very hard: (0,0.3].

In summary, we model typing performance by building a distribution of RDA scores from real training data under different artifact conditions. This model is then simulated typing 10 sentences 15 times to compare the performances of robust and non-robust classifiers. We report our results in terms of typing accuracy and duration (total seconds per word completion), see Figures 2 and 3. For reference, we include the area under the curve (AUC) values for each subject under all artifact conditions in 1.

Table 1. AUC values

	Subject 1	Subject 2	Subject 3	Subject 4
No-Artifact	.7644	.8298	.6488	.8103
Jaw Movement	.6079	.8026	.6370	.6527
Smile	.7105	.8423	.6506	.7023
Eye Blink	.6561	.7641	.4710	.7373

4.1 Typing Accuracy

As can be noted in Figure 2, typing accuracy changes dramatically between subjects. In the simulation, as with other trials we've performed, subject 3 struggles to produce accurate classifications. Additionally, we note that robust classification consistently outperforms non-robust methods. The performance advantage of our method is correlated to the magnitude of the difference in AUC between the control and artifact classes. In other words, the stronger the drop in AUC when an artifact is introduced (Table 1), the greater the performance benefit of using robust fusion.

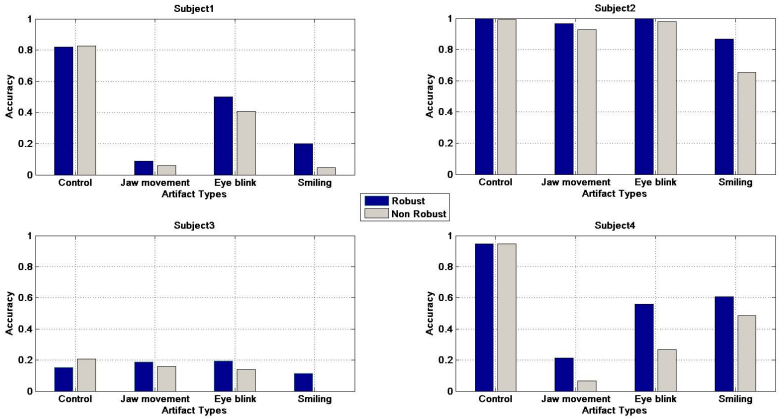


Fig. 2. Accuracy vs artifact type

4.2 Typing Duration

From Figure 3, we immediately notice that the robust case typically types faster than the non-robust case. Additionally, considering the AUC values from Table 1 and the results from Figure 3, we notice that higher AUC values offer quicker typing performance. Both these effects share a common motivation. The typing system repeats sequences until a sufficiently high confidence threshold is reached. Accurate typing, because of robust methods or high user AUC, will yield fast typing.

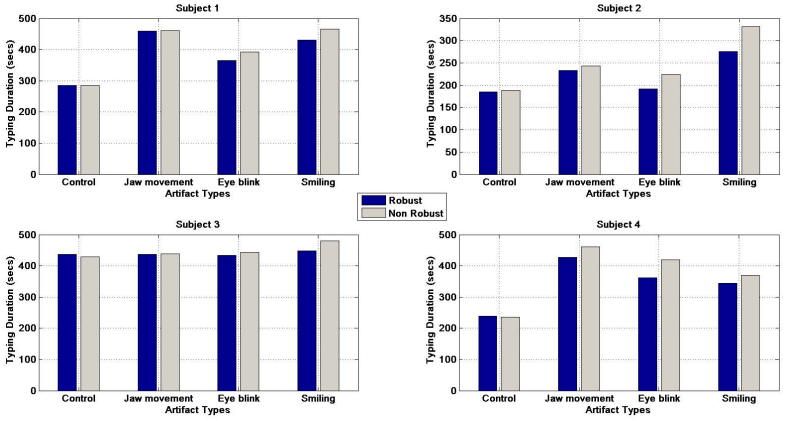


Fig. 3. Typing duration vs artifact type

5 Conclusions

We designed a robust classification method for ERP detection in a BCI typing paradigm. We tested the proposed method on the RSVP Keyboard^{TM}, which is an in-house BCI typing system. Considering the possible changes in the EEG data, we developed a mixture density model for class conditional EEG evidence to use in the fusion with n-gram language model. To compare the robust and non-robust classification methods, using pre-recorded calibration data, we simulated the performance of four subjects typing 10 sentences 15 times and reported results on accuracy and speed of their typing.

Each of our simulations was run under a single artifact class (rather than a mixture of multiple classes). When implemented with a true mixture of artifact classes we observed nearly identical results between robust and non-robust methods. We suggest this is due to the ability of our classifier to accumulate additional EEG evidence when an input doesn't reach the confidence threshold. In the true mixture case, where artifacts aren't very frequent, the classifier is bound to receive useful information during the following sequences. For further analysis, we are interested in examining the cause of mis-classified symbols. We hypothesize that the risk in artifacts is not readily seen in their prior distribution as artifacts frequently occur in bursts during operation. Future artifact class models which are conditioned on previous artifact classes, allowing for bursts of artifacts to occur while still keeping artifact priors at reasonable levels, will be studied.

References

1. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Muller, K.R.: Optimizing spatial filters for robust eeg single-trial analysis. IEEE Signal Processing Magazine 25(1), 41–56 (2008)
2. Nezamfar, H., Orhan, U., Erdogmus, D., Hild, K., Purwar, S., Oken, B., Fried-Oken, M.: On visually evoked potentials in eeg induced by multiple pseudorandom binary sequences for brain computer interface design. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2044–2047 (2011)
3. Krusienski, D.J., Sellers, E.W., McFarland, D.J., Vaughan, T.M., Wolpaw, J.R.: Toward enhanced P300 speller performance. Journal of Neuroscience Methods 167(1), 15–21 (2008)
4. Treder, M., Blankertz, B. (C)overt attention and visual speller design in an erp-based brain-computer interface. Behavioral and Brain Functions 6(1), 28 (2010)
5. Friedman, J.H.: Regularized discriminant analysis. Journal of the American Statistical Association 84(405), 165–175 (1989)
6. Roark, B., de Villiers, J., Gibbons, C., Fried-Oken, M.: Scanning methods and language modeling for binary switch typing. In: The NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, p. 2836. Association for Computational Linguistics (2010)

Real-Time Vigilance Estimation Using Mobile Wireless Mindo EEG Device with Spring-Loaded Sensors

Li-Wei Ko[1,2], Chun-Hsiang Chuang[2,3], Chih-Sheng Huang[2,3], Yen-Hsuan Chen[2,3], Shao-Wei Lu[2,3], Lun-De Liao[2,3], Wan-Ting Chang[2,3], and Chin-Teng Lin[2,3*]

[1] Department of Biological Science and Technology, National Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan
[2] Brain Research Center, National Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan
[3] Institute of Electrical Control Engineering, National Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan
{Lwko,swlucifer,ctlin}@mail.nctu.edu.tw, chchuang@ieee.org, {chih.sheng.huang821,gs336.tw,claire86110}@gmail.com, terry7886.ece96@g2.nctu.edu.tw

Abstract. Monitoring the neurophysiological activities of human brain dynamics in an operational environment poses a severe measurement challenge using current laboratory-oriented biosensor technology. The goal of this research is to design, develop and test the wearable and wireless dry-electrode EEG human-computer interface (HCI) that can allow assessment of brain activities of participants actively performing ordinary tasks in natural body positions and situations within a real operational environment. Its implications in HCI were demonstrated through a sample application: vigilance-state prediction of participants performing a realistic sustained-attention driving task. Besides, this study further developed an online signal processing for extracting EEG features and assessing cognitive performance. We demonstrated the feasibility of using dry EEG sensors and miniaturized supporting hardware/software to continuously collect EEG data recorded from hairy sites (i.e., occipital region) in a realistic VR-based dynamic driving simulator.

Keywords: Drowsy driving, Wireless and dry EEG device, Mindo, Human-computer interface.

1 Introduction

Conventional wet electrodes are commonly used to measure EEG signals [1], and they provide excellent EEG signals with the proper skin preparation and conductive gel application. However, a series of skin preparation procedures for applying the wet electrodes is always required and usually creates trouble for users [2]. Further, the

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 450–458, 2013.

signal quality may degrade over time as the skin regenerates and the conductive gel drives. Recently, measuring the EEG signals using the dry EEG sensors have become available — foam-based sensors [3] for example. However, there are still some reminding issues that need to improve. For instance, a small part of subjects would have allergy issues, even the materials were modified as a bio-compatible one. During an hour-long cognitive experiment, the sensors have attached on the skin surface for a long time, resulting in that the subjects sweating on the sensors surface, which consequently cause some unknown reactions on the skin. In addition, the foam-based sensors are only useful for non-hairy sites such as the forehead. The EEG acquisition from the hairy sites is still a challenge.

To overcome these drawbacks, a new dry-contact EEG device [2, 4] with spring-loaded sensors [5] was used for potential operation in the presence or absence of hair and without any skin preparation or conductive gel usage. Significantly, the flexibility of the proposed dry EEG sensor is effective in tightly contacting the scalp surface and providing clear EEG signals without any skin preparation or conductive gel usage.

This work demonstrated the feasibility of the wireless and mobile EEG device using spring-loaded sensors through a typical human-computer interface application: monitoring human cognitive states in a realistic sustained-attention driving task [6]. The online HCI system comprised three major modules: EEG receiving, signal processing, and display modules. In signal processing, we designed an effective algorithm to allow a JAVA platform to implement on-line signal processing. Taken together, this integrated neuroergonomic system capable of measuring and processing concurrent neural, behavioral, psychophysiological, environmental, and system operational data could allow continuous estimation of subjects' cognitive state to design and operate systems that maximize operator cognitive capacity as well as overall human/system performance.

2 Materials and Methods

2.1 Experimental Task and Subjects

A sustained-attention driving task (event-related lane-departure paradigm [6]) was implemented in a virtual-reality (VR) driving simulator [7]. The VR driving environment consists of a 360-degree surrounding vision that simulates a nighttime driving on an uncrowded highway. The used paradigm was to try to induce subjects' drowsiness and obtain their drowsy patterns, including EEG signals and behaviors. During a 1.5 hr. experiment, participants were instructed to compensate for the trajectory error as soon as possible while they detect the deviation event. The deviation event randomly occurred. The duration time in response to the deviation event, denoted as the response time (RT), was used as an indicator to evaluate subject's vigilance level and also used to label EEG trial.

Seven volunteers participated in this experiment. Each participant was required to have a lunch at noon and the task would start at 1:30 PM. As shown in Fig. 1, each participant wearing the Mindo EEG device sat inside the vehicle simulator, and controlled the simulator by using the steering wheel.

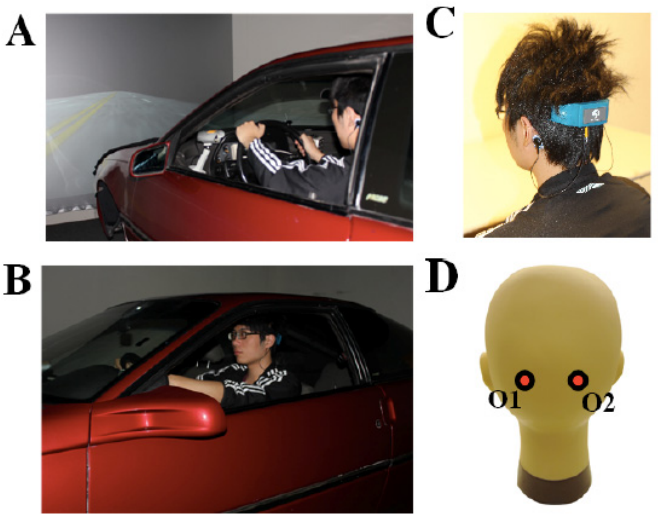


Fig. 1. Experimental task and EEG device. (A) An immersive driving environment. (B) A participant wears the Mindo device in the driving simulator. (C) The Mindo collects EEG signal from (D) the occipital area (i.e., O1 and O2 channel).

2.2 Systematic Diagram

Fig. 1 shows the scheme of the proposed driver fatigue prediction system that includes the EEG data collection (left panel), system construction (middle panel), and real-time EEG analysis (right panel). After database collection, the training process was implemented to construct the vigilance prediction system. Previous EEG studies [8-10] showed the vigilance state was significantly correlated with the power spectrum of EEG dynamics. For instance, the spectral power of the alpha activity in the drowsiness state was stronger than that in the alert state. In this study, the fast Fourier transformation (FFT) was used to extract EEG features. To obtain an accurate estimation, we proposed a weighted algorithm for on-line time-frequency analysis. The detail is described in Section 2.5.

As our previous work [11], the support vector regression [12] was applied to construct the core algorithm of the prediction system, in which the independent variables were the power spectrum array (1-30 Hz) and dependent variables were the RTs. All of the algorithms and signal processing methods were implemented in a JAVA interface to integrate the software and hardware into a complete system. The developed system is an easy and effective way for monitoring the driver physiological state in real-time.

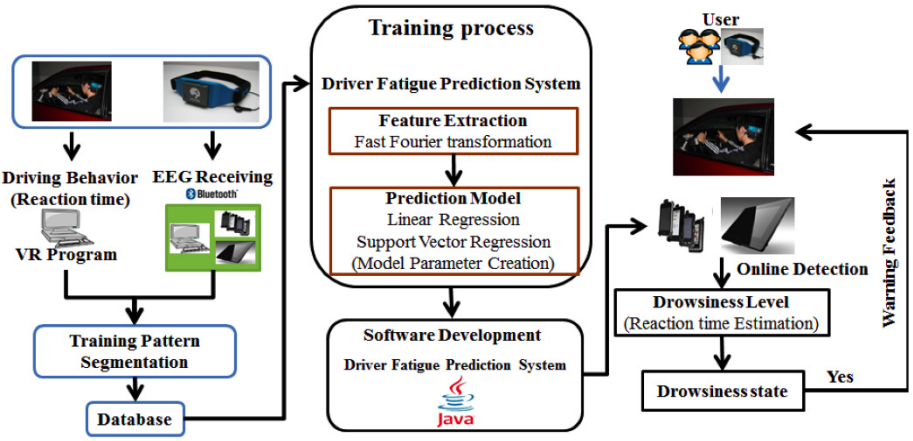


Fig. 2. Systematic diagram for constructing a driver's vigilance prediction system

2.3 EEG and Behavior Recorded by Mindo

Figs. 3 (A) and (B) show the Mindo system with one ground channel, one reference channel, and four EEG channels. The 4-channel EEG signals (in μV) and 1-channel behavioral response (in pt.) simultaneously recorded by the acquisition software. The server with Bluetooth module wirelessly received EEG signals from the Mindo device at sampling rate of 256 Hz. The analogue signals were converted into the digital form with a 16-bit resolution in the range of -1.5V~1.5V. In addition, the server also received the data via RS-232 compatible serial port from the client which runs the VR program and recorded the behavioral response. This data stream with an 8-bit digital resolution including the vehicle trajectory (0-240), deviation onset (251/252 for left and right side of the deviation), response onset (253), and response offset (254) was synchronized with the EEG data for further event-related analysis.

2.4 Spring-Loaded Sensors

Fig. 3 (C) shows the spring-load sensor [5] designing by the conductive metal thimble and spring material. The spring-load sensor, which is a dry sensor and contacts with scalp directly, overcomes the problem from thick hair and long-term monitoring. Therefore, the spring-load sensor is more suitable and convenient than traditional conductive gel in real applications. The dry EEG sensors were designed to contact the scalp surface with 17 spring contact probes. Each probe was designed to include a probe head, plunger, spring, and barrel. The 17 probes were inserted into a flexible substrate using a one-time forming process via an established injection molding procedure. With 17 spring contact probes, the flexible substrate allows for high geometric conformity between the sensor and the irregular scalp surface to maintain low

skin-sensor interface impedance. Additionally, the flexible substrate also initiates a sensor buffer effect, eliminating pain when force is applied. The used dry EEG sensor was reliable in measuring EEG signals without any skin preparation or conductive gel usage, as compared with the conventional wet electrodes.

Fig. 3. Mindo with spring-load sensors

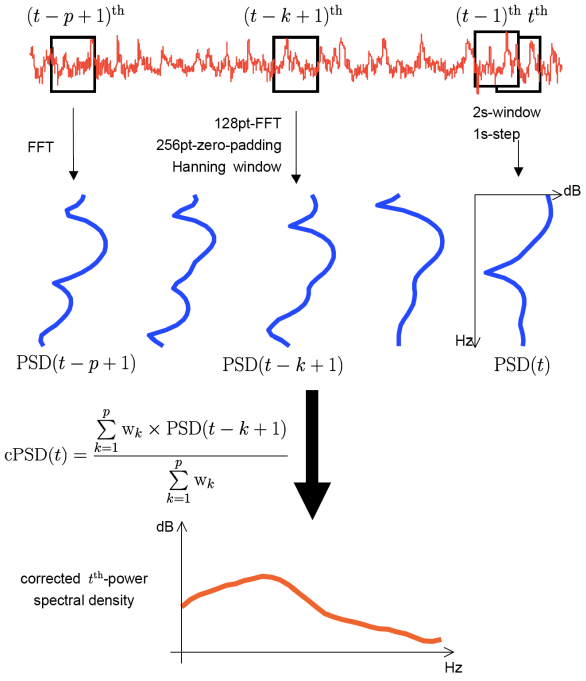


Fig. 4. Weighted spectral power estimation

2.5 Weighted Time-Frequency Analysis

The EEG signal was successively feed into weighted time-frequency analysis before applying support vector regression. As shown in Fig. 4, power spectral density (PSD) of the *t*-th EEG trial (a 2-s EEG signal) were the weighted average of spectral powers which were calculated from {*t*-*p*+*1*}-th, …, {*t*-*k*-*1*}-th, …, *t*-th EEG trials, where k ≤ *p*. Windowed 128-point epochs were extended to 256 points by zero-padding. The obtained EEG power spectra were further converted to a logarithmic scale prior to further analysis. Then, a weighted-averaging filter was used on all the PSDs to further obtain a smoothing a PSD estimation. In practice, we multiplied each PSD by a weighted coefficient, $\{w_k | k = 1,2, \dots, p\}$, independently, in which w_k increased as k decreased. In this study, $w = 1,2, \dots, 20$ which means that there are 20 windows ($p = 20$) for each PSD estimation. As shown in Fig. 4, compared to an unprocessed PSD (the blue traces), we can obtain a more accurate PSD estimation (the red trace) by using this algorithm.

3 Experimental Results

Table 1 shows the prediction results for each subject. The performances were compared by calculating the correlation coefficient and the root mean square error (RMSE) between observed RT and predicted RT. As can be seen, the performance can reach the correlation coefficient of 0.9471, 0.7882, 0.8475, 0.8920, 0.9370, 0.9738, and 0.9559 which means that the similarity between the observed RT and the predicted RT are very high. In terms of RMSE, most of the errors are lower than 0.1 second.

Table 1. Results of the prediction within subject validation

	Subjects	Correlation coefficient	RMSE (unit: millisecond)
Within subject validation	S01	0.9471	50.6737
	S02	0.7882	141.8695
	S03	0.8475	80.3044
	S04	0.8920	80.6624
	S05	0.9370	109.6332
	S06	0.9738	60.3334
	S07	0.9559	70.1674

Table 2 shows the prediction results using leave-one-subject-out cross validation. At each step of cross validation, the support vector regression is trained on EEG data from six subjects and tested on the remaining subject. This procedure repeats for all subjects being a test dataset. All the parameters of the system were calculated from the training data and applied to the testing data. As can be seen, in most of the cases (i.e., S03, S04, S05, S06, and S07) the system still can obtain a robust prediction result (correlation coefficient > 0.8).

Table 2. Results of the prediction using leave-one-subject-out cross validation

	Subjects	Correlation coefficient	RMSE (unit: millisecond)
Leave-one-subjec-out validation	S01	0.5768	548.0701
	S02	0.4799	839.2128
	S03	0.8886	174.6283
	S04	0.8378	369.1276
	S05	0.8932	552.2017
	S06	0.8848	607.8501
	S07	0.8493	1009.2067

Fig. 5 shows the result of the real-time vigilance prediction system. The black trace indicates the vehicle trajectory. As observed in Fig. 5, we could find that the driving errors became large as time went by. The blue trace is the predicted RT. The result showed that the predicted RT had an increasing trend that co-varied with the driving errors.

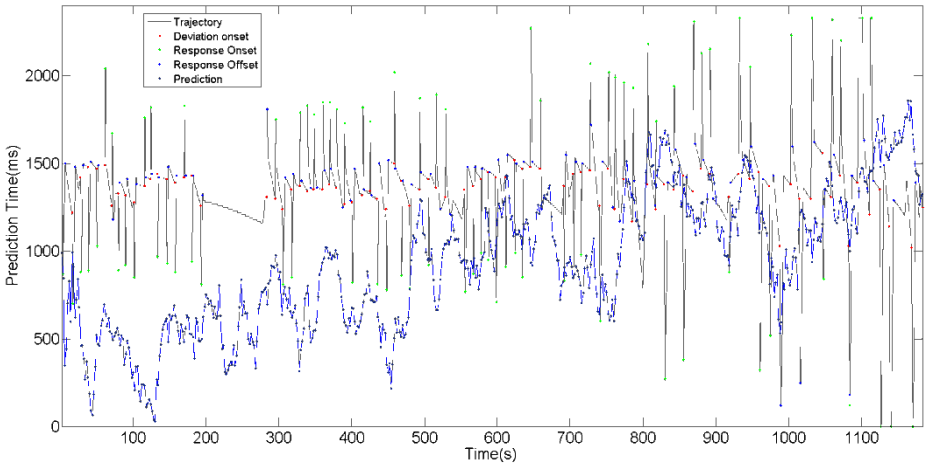


Fig. 5. The result of the real-time vigilance prediction system

4 Discussions and Conclusions

This study proposed a novel HCI system that can continuously monitor driver's vigilance state in real-time. Our empirical results showed that the efficacy of EEG signal acquisition was much more effective and easier to collect human brain activities in an operational environment. The remaining issue is to develop an algorithm to automatically remove artifacts for improving the system performance.

In conclusions, this study incorporated a rich interconnection between previously established, conventional laboratory-derived theoretical bases, novel EEG device and testing within fairly complex scenarios and environments. Furthermore, the dry EEG devices with spring-loaded sensors promote more HCI applications in natural environments.

Acknowledgement. This work was supported in part by the UST-UCSD International Center of Excellence in Advanced Bio-engineering sponsored by the Taiwan National Science Council I-RiCE Program under Grant Number: NSC-101-2911-I-009-101, in part by the Aiming for the Top University Plan of National Chiao Tung University, the Ministry of Education, Taiwan, under Contract 102W963, and in part by the National Science Council, Taiwan, under Contract 100-2628-E-009-027-MY3. Research was also sponsored in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation heron.

References

1. Gevins, A., Le, J., Martin, N.K., Brickett, P., Desmond, J., Reutter, B.: High resolution EEG: 124-channel recording, spatial deblurring and MRI integration methods. Electroencephalogr Clin Neurophysiol 90(5), 337–358 (1994)
2. Liao, L.-D., Lin, C.-T., McDowell, K., Wickenden, A.E., Gramann, K., Jung, T.-P., Ko, L.-W., Chang, J.-Y.: Biosensor technologies for augmented brain-computer interfaces in the next decades. Proceedings of the IEEE 100(Special Centennial Issue), 1553–1566 (2012)
3. Lin, C.-T., Liao, L.-D., Liu, Y.-H., Wang, I.-J., Lin, B.-S., Chang, J.-Y.: Novel dry polymer foam electrodes for long-term EEG measurement. IEEE Transactions on Biomedical Engineering 58(5), 1200–1207 (2011)
4. Liao, L.-D., Chen, C.-Y., Wang, I.-J., Chen, S.-F., Li, S.-Y., Chen, B.-W., Chang, J.-Y., Lin, C.-T.: Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors. Journal of Neuroengineering and Rehabilitation 9(5), 1–11 (2012)
5. Liao, L.-D., Wang, I.-J., Chen, S.-F., Chang, J.-Y., Lin, C.-T.: Design, Fabrication and Experimental Validation of a Novel Dry-Contact Sensor for Measuring Electroencephalography Signals without Skin Preparation. Sensors 11(6), 5819–5834 (2011)
6. Huang, R.-S., Jung, T.-P., Makeig, S.: Tonic changes in EEG power spectra during simulated driving. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) Augmented Cognition, HCII 2009. LNCS, vol. 5638, pp. 394–403. Springer, Heidelberg (2009)
7. Lin, C.-T., Chuang, C.-H., Wang, Y.-K., Tsai, S.-F., Chiu, T.-C., Ko, L.-W.: Neurocognitive characteristics of the driver: A review on drowsiness, distraction, navigation, and motion sickness. Journal of Neuroscience and Neuroengineering 1(1), 61–81 (2012)
8. Cantero, J.L., Atienza, M., Salas, R.M.: Human alpha oscillations in wakefulness, drowsiness period, and REM sleep: different electroencephalographic phenomena within the alpha band. Clinical Neurophysiology 32(1), 54–71 (2002)

9. Lin, C.-T., Wu, R.-C., Liang, S.-F., Chao, W.-H., Chen, Y.-J., Jung, T.-P.: EEG-based drowsiness estimation for safety driving using independent component analysis. IEEE Transactions on Circuits and Systems I: Regular Papers 52(12), 2726–2738 (2005)
10. Makeig, S., Jung, T.-P.: Tonic, phasic, and transient EEG correlates of auditory awareness in drowsiness. Cognitive Brain Research 4(1), 15–25 (1996)
11. Lin, F.-C., Ko, L.-W., Chuang, C.-H., Su, T.-P., Lin, C.-T.: Generalized EEG-based drowsiness prediction system by using a self-organizing neural fuzzy system. IEEE Transactions on Circuits and Systems I: Regular papers 59(9), 2044–2055 (2012)
12. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing 14(3), 199–222 (2004)

Relationship Analysis between Subjective Evaluation and NIRS-Based Index on Video Content

Shinsuke Mitsui[1], Atsushi Maki[2], and Toshikazu Kato[3]

[1] Industrial and Systems Engineering, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan
a12.ep8r@g.chuo-u.ac.jp
[2] Hitachi, Ltd, 1-18-13, Soto-Kanda, Chiyoda-ku, Tokyo, 101-8608, Japan
atsushi.maki.nn@hitachi.com
[3] Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan
kato@indsys.chuo-u.ac.jp

Abstract. Brain activities have been investigated, and various functions of brain have been revealed recently. In our experiment, decrease of oxy-Hb change at frontal cortex was observed while subjects were watching video contents. Also, the degrees of decreases were different among the subjective evaluations about impression against the video contents. Revealing the cause of the decrease has the possibility to evaluate video content objectively. In this paper we discuss the relationship between subjective evaluation and brain activity on video content.

Keywords: Frontal cortex, Near-infrared spectroscopy, Subjective evaluation, Video content.

1 Introduction

A traditional evaluation method for video contents has been based on customers' reviews, and the numbers of them. However, the method has some problems on reliability and, needs to embrace scientific approaches regarding to brain activity [1]. Thus, in this study, we focused on a brain activity, and analyzed the relationship between subjective evaluations of video contents and brain activity.

We observed brain activity at frontal cortex during watching video contents, as frontal cortex was related to higher-level cognitive function. We used Near-infrared spectroscopy (NIRS), as it was non-invasive and restraint-free, less noise than electroencephalography (EEG), and higher temporal resolution than functional magnetic resonance imaging (fMRI) [2].

Nine healthy right-handed subjects watched six videos (two excellent videos, two average videos and two poor videos). We used NIRS to measure oxygenated hemoglobin (oxy-Hb) change in frontal cortex. Consequently we found a decreased oxy-Hb change during watching video (TV Commercial). In addition, there are significant differences among excellent videos, average videos, and poor videos in oxy-Hb change.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 459–466, 2013.

These results suggest that the decrease in oxy-Hb change is related to subjective evaluations of video content.

2 Subjective Evaluation of Video Contents for NIRS Experiment

2.1 Subjects

Thirty healthy Japanese adults (22 men and 8 women, aged 21 to 25) participated in first subjective evaluation. Subjects were divided into three groups. In second subjective evaluation after the first evaluation, other nine healthy adults (all were men, aged 21 to 24 and right-handed), who were differ from the thirty subjects, participated to confirm the result of the subjective evaluation.

2.2 Procedures

In the first subjective evaluation, each thirty subject watched sixteen videos, and answered a questionnaire to gather the evaluation about overall impression, background music, story, persona, company, and product by the 5-point rating scale: Very Poor (1), Poor (2), Average (3), Good (4), Excellent (5). In this paper, we used only the evaluation about overall impression in analysis. We used 30-second TV Commercials as video contents. The commercial were related to products, such as home electronics, foods, etc., and companies themselves. The orders of the sixteen videos were random for each group. Their ratings were averaged for each video, and sixteen videos were ranked by the average score.

Also, we picked up the top two videos as excellent, the worst two videos as poor and two videos which are near 3.00 point as average.

In the second subjective evaluation, each nine other subject watched the six videos (the excellent, average, and poor videos) and answered the same questionnaire to confirm the previous result of ranking.

2.3 Result

Table1 and Fig.1 showed the result of subjective evaluation of sixteen videos in thirty subjects, also Table2 showed six videos; two excellent videos, two average videos, and two poor videos.

Table3 and Fig.2 showed the result of subjective evaluation of the six videos in nine subjects.

Compared with Table2 and Table3, the ranking of the six videos was almost the same between thirty subjects and the other nine subjects. Thus, these videos were elected accurately as excellent, average, and poor among the adults who aged 21 to 25. These six videos were used in NIRS experiment.

Table 1. Sixteen videos (TV Commercials) list. The videos were sorted in descending order of the average score about overall impression by 5-point scale.

CM No.	Year	Company name	Average score
1	1980	FUJIFILM Corporation	4.39
2	1974	Panasonic Corporation	4.13
3	1976	Meiji Holdings Co., Ltd.	3.96
4	1982	Tokio Marine & Nichido Fire Insurance Co., Ltd.	3.91
5	1987	Sony Marketing (Japan) Inc.	3.91
6	1990	Sony Marketing (Japan) Inc.	3.83
7	1970	Panasonic Corporation	3.74
8	1976	Shiseido Company, Limited	3.61
9	1978	Ryukakusan Co., Ltd.	3.48
10	1978	LOTTE Co., Ltd.	3.43
11	1973	Sony Marketing (Japan) Inc.	3.09
12	1981	Lion Corporation	3.09
13	1977	Panasonic Corporation	3.04
14	1969	PILOT CORPORATION	2.96
15	1985	Shiseido Company, Limited	2.74
16	1980	Yamaha Motor Co., Ltd.	2.61

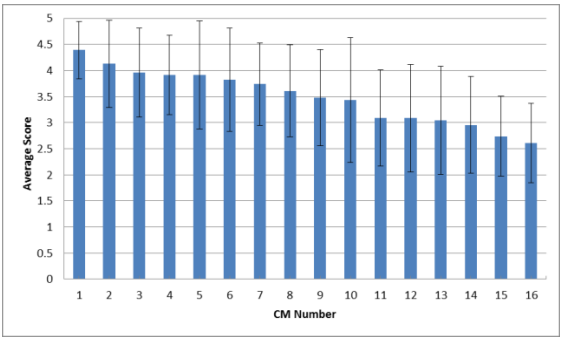


Fig. 1. Average score and standard deviation of sixteen videos. The standard deviations showed that there were differences among individuals.

Table 2. Videos; excellent, average and poor videos in the experiment in thirty subjects

CM No.	Year	Company name	average score	evaluation
1	1980	FUJIFILM Corporation	4.39	Excellent
2	1974	Panasonic Corporation	4.13	Excellent
13	1977	Panasonic Corporation	3.04	Average
14	1969	PILOT CORPORATION	2.96	Average
15	1985	Shiseido Company, Limited	2.74	Poor
16	1980	Yamaha Motor Co., Ltd.	2.61	Poor

Table 3. Videos; excellent, average, and poor videos in the experiment in other nine subjects. The ranking was almost the same as the result of previous ranking in thirty subjects. Only the rank of No.14 (PILOT CORPORATION, 1969) and the rank of No.15 (Shiseido Company, 1985) were reversed. We used these six videos for NIRS experiment.

CM No.	Year	Company name	Average score	Evaluation
1	1980	FUJIFILM Corporation	4.67	Excellent
2	1974	Panasonic Corporation	4.33	Excellent
13	1977	Panasonic Corporation	2.67	Average
15	1985	Shiseido Company, Limited	2.56	Average
14	1969	PILOT CORPORATION	2.22	Poor
16	1980	Yamaha Motor Co., Ltd.	2	Poor

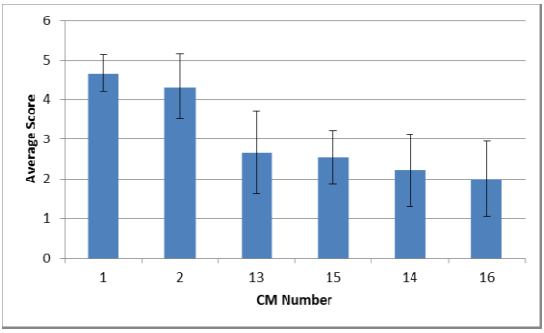


Fig. 2. Average score and standard deviation of six videos. There was an obvious difference between excellent(CM No.1 and 2) and poor videos(CM No.14 and 16).

3 NIRS Experiment during Watching Videos

3.1 Subjects

Nine healthy Japanese adults (the same 9 subjects in the previous subjective evaluation of confirmation) participated in this NIRS experiment.

3.2 Procedure

This NIRS experiment was conducted concurrently while the nine subjects were evaluating 6 videos. Each subject was seated in front of a table on which 17-inch display was placed. Changes in the concentration of oxy-Hb were measured at 22 channels with an ETG-4000(HITACHI Medical Corporation, Japan) during the watching of six videos in Table3. We measured frontal cortex area according to the international 10-20 system in electroencephalography (Fig.3, Fig.4) [3] [4]. Each subject watched and evaluated the overall impression of the six videos one by one using the 5-point rating. The 6 videos were presented on the display, and in random for each participant to avoid order effect. A block paradigm was used in a design of experiment for watching six times repetition of video [5]. A condition was 30-second period of video task (CM Task) with a 40-second period of rest. The rest consisted of 25-second rest time, 5-second evaluation time for overall impression by hands, and 10-second rest time (Fig.5). This condition was repeated 6 times for each participant.

3.3 Data Analysis

Oxy-Hb changes of two excellent, two average, and two poor videos were averaged and the grand averaged waveforms were made. Also grand averaged waveforms were compared by t-test between excellent and average videos, poor and average videos, and excellent and poor videos.

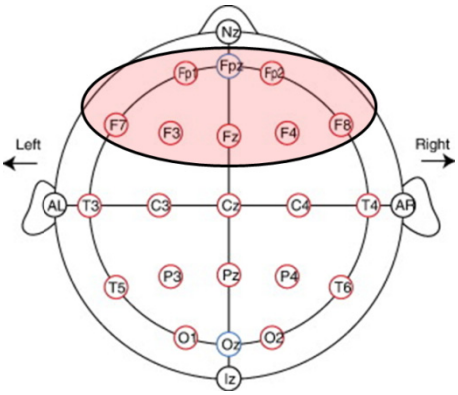


Fig. 3. Names and positions of international 10-20 system in this study [4]. We measured frontal cortex area which was marked.

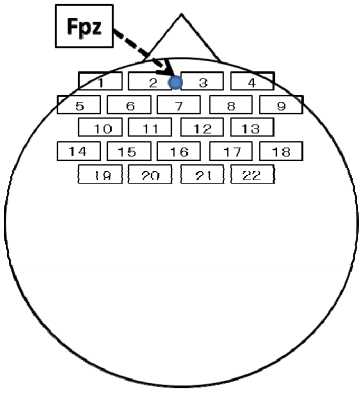


Fig. 4. The location of 22 channels. We located the middle of channel 2 and channel 3 at Fpz.

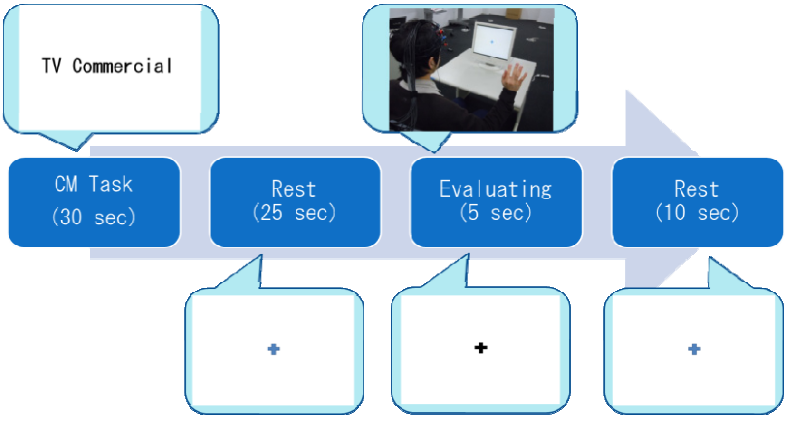


Fig. 5. Design of experiment. This cycle was repeated 6 times for a subject.

3.4 Results of NIRS Experiment

According to the Fig.6, a decrease of oxy-Hb change was observed in each video. Also, the degree of decrease was different from each other.

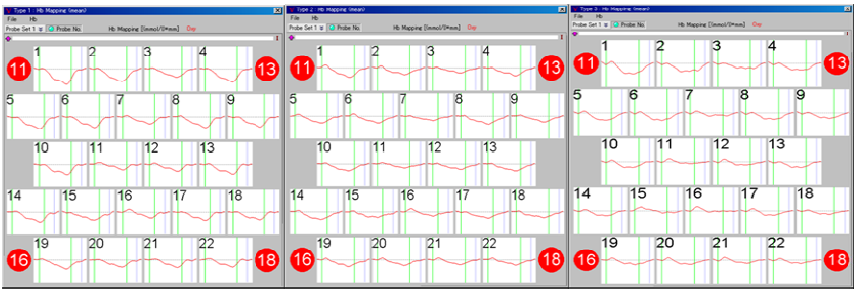


Fig. 6. Grand averaged waveforms (left; Excellent videos, middle; Average videos, right; Poor videos). The number of subjects is nine.

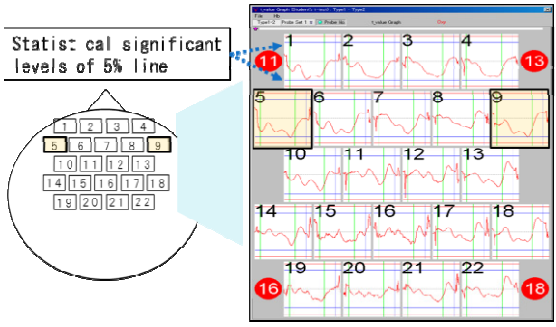


Fig. 7. t-value graphs of oxy-Hb comparison between excellent and average videos. Excellent videos were significantly smaller than average videos in the channels which were marked by a heavy line.

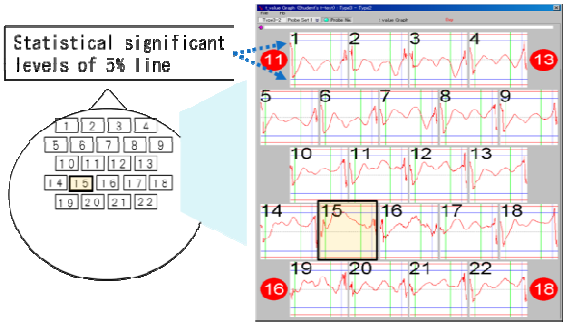


Fig. 8. t-value graphs of oxy-Hb comparison between poor and average videos. Poor videos were significantly larger than average videos in the channels which were marked by a heavy line.

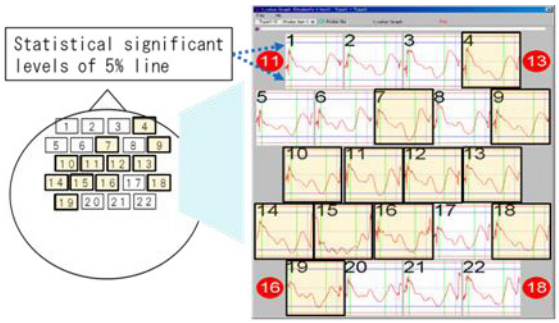


Fig. 9. t-value graphs of oxy-Hb comparison between excellent and poor videos. Excellent videos were significantly smaller than poor videos in the channels which were marked by a heavy line.

The result of the t-test for oxy-Hb changes showed some significant differences. Excellent videos were smaller than average videos in two channels (ch5 and ch9) (Fig.7). Poor videos were bigger than average videos in 1 channel (ch15) (Fig.8). Excellent videos were smaller than poor videos in 13 channels (ch4, 7, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, and 21) (Fig.9).

4 Conclusion and Discussion

We picked out excellent, average, and poor videos (TV commercials) properly about overall impression by using 5-point rating scale questionnaire. Also, we observed decrease of oxy-Hb changes in frontal cortex while subjects were watching the videos, and found that degrees of decreases were different among the evaluation of overall impressions. These results suggest that NIRS at frontal cortex can detect differences of subjective evaluations of video content. In addition, the results also suggest that we can measure the degree of the subjective evaluation by investigating change of oxy-Hb at the channels that showed significant differences.

Recently some researches have indicated the decrease of oxy-Hb change in various brain areas, and decrease of oxy-Hb change has been defined as brain deactivation [6] [7]. Also, some processes of deactivation at frontal cortex have been suggested. Two of them can be related to our decrease of oxy-Hb change. One is deactivation caused by engaging in goal-directed actions, such as detecting targets, because our task in which subjects are watching TV commercials can be as goal-directed action. Another is deactivation caused by conducting task that involve externally focused attention, because our TV commercials were so old that subjects could accept the TV commercials as external stimulation. In future research, we need to find out the cause of deactivation of our study along the two processes.

Acknowledgements. This work was partially supported by JSPS KAKENHI grants, "Effective Modeling of Multimodal KANSEI Perception Processes and its Application to Environment Management" (No. 24650110), "Robotics modeling of

diversity of multiple KANSEI and situation understanding in real space" (No. 19100004) and TISE Research Grant of Chuo University, "KANSEI Robotics Environment".

References

1. Laura, A. F., De Vico Fallani, F., Cincotti, D., Mattia, L., Bianchi, M.G., Marciani, S., Salinari, A., Colosimo, A., Tocci, R., Soranzo, F., Babiloni: Neural basis for brain responses to TV Commercials: A high-resolution EEG study. IEEE Transactionson Neural Systemand Rehabilitation Engineering 16(6) (2008)
2. Kono, T., Matsuo, K., Tsunashima, K., Kasai, K., Takizawa, R., Rogers, M.A., Yamasue, H., Yano, T., Taketani, Y., Kato, N.: Multiple-time replicability of near-infrared spectroscopy recording during prefrontal activation task in healthy men. Neurosci. Res. 57, 504–512 (2007)
3. Okamoto, M., Dan, H., Sakamoto, K., Takeo, K., Shimizu, K., Kohno, S., Oda, I., Isobe, S., Suzuki, T., Kohyama, K., Dan, I.: Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10-20 system oriented for transcranial functional brain mapping. NeuroImage 21(1), 99–111 (2003)
4. Jurcak, V., Okamoto, M., Singh, A., Dan, I.: Virtual 10-20 measurement on MR images for inter-modal linking of transcranial and tomographic neuroimaging methods. NeuroImage 26(4), 1184–1192 (2005)
5. Ye, J.C., Tak, S., Jang, K.E., Jung, J., Jang, J.: NIRS-SPM: Statistical parametric for near-infrared spectroscopy. NeuroImage 44, 428–447 (2008)
6. Shimada, S.: Deactivation in the sensorimotor area during observation of a human agent performing robotic actions. Brain and Cognition 72(2010), 394–399 (2010)
7. Gusnard, D.A., Raichle, M.E.: Searching for a Baseline: Functionalimaging and the resting human brain. Nature Reviews Neuroscience 2, 685–694 (2001)

Towards Evaluating Computational Models of Intuitive Decision Making with fMRI Data

James Niehaus, Victoria Romero, and Avi Pfeffer

Charles River Analytics, Inc., Cambridge, MA
{jniehaus,vromero,apfeffer}@cra.com

Abstract. A vast array of everyday tasks require individuals to use intuition to make decisions and act effectively, including civilian and military professional tasks such as those undertaken by firefighters, police, search and rescue, small unit leaders, and information analysts. To better understand and train intuitive decision making (IDM), we envision future training systems will represent IDM through computational models and use these models to guide IDM learning. This paper presents the first steps to the problem of validating computational models of IDM. To test if these models correlate with human performance, we examine methods to analyze functional magnetic resonance imaging (fMRI) data of human participants performing intuitive tasks. In particular, we examine the use of a new deep learning representation called sum-product networks to perform model-based fMRI analysis. Sum-product networks have been shown to be simpler, faster, and more effective than previous deep learning approaches, making them ideal candidates for this computationally demanding analysis.

Keywords: intuition, intuitive decision making, deep learning, sum-product network, functional magnetic resonance imaging, model-based fMRI.

1 Introduction

A vast array of everyday tasks require individuals to use intuition to make decisions and act effectively, including civilian and military professional tasks such as those undertaken by firefighters, police, search and rescue, small unit leaders, and information analysts. Currently, intuition is developed only incidentally after years of training and on-the-job experience, costing time, money, and potentially lives as trainees are unable to perform to the intuitive decision making (IDM) needs of their positions.

To address this need, research must be performed to establish a scientific and technical basis for IDM. Specifically, major research areas include:

1. **Characterizing intuitive decision making and implicit learning across neural, cognitive, and behavioral levels of representation:** We define intuition as a rapid, non-conscious mental process that may operate with limited or uncertain information to produce a judgment or response [1,2]. An accurate characterization

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 467–473, 2013.

will incorporate the neural basis for intuition, the cognitive components of intuitive decision making, and the behavioral outcomes of intuitive action.

2. **Representing intuitive decision making through computational models:** The computational models must adapt to new information and use context and knowledge to make intuitive decisions. To provide training recommendations, the computational models must capture the differences in intuitive decision making as a result of learning and individual differences.

In addition to addressing each of the major areas above, a program of study must support solutions that are validated against human-performance data. In this paper, we focus on a technique to validate computational models of intuitive decision making (IDM) with functional magnetic resonance imaging (fMRI) human-performance data. The basic approach is to give human participants an IDM task within a fMRI machine and compare the recorded fMRI and behavioral data with the predictions of the computational model. This process is known as a model-based fMRI analysis [3]. We extend existing model-based fMRI analyses by examining a deep learning method, called sum-product networks [4], to constructing temporal models of relationships among stimuli, computational models, fMRI data, and recorded behavior.

2 Related Work

2.1 Computational Models of Intuitive Decision Making

There have been two tasks consistently used to explore the neural correlates of intuitive decision making (IDM). First, more than a dozen studies have used the serial reaction time task (SRTT) [5] which involves pressing buttons as quickly as possible to indicate the position of a target on the screen. The second task used to examine IDM is the artificial grammar task (AGL) [6]. The AGL involves exposing participants to letter strings (e.g., "TQSLV") that are all generated based on a hidden set of rules derived from a Markovian grammar chain. Computationally, these two tasks can represented as learning important features from of phenomena and grouping those phenomena into categories that are functionally similar based on those features.

Dirichlet process (DP) mixture model [7,8] provides a prior over partitions of a set of observations into groups, each with its own distribution over parameters. Dirichlet process mixture models can be extended to become hierarchical DP (HDP) mixture models [9]. These models provide a simple recipe for representing the densities associated with multiple categories simultaneously. Recent research has shown how HDP mixture models can unify classical prototype and exemplar models of human categorization, adaptively transitioning from prototype-like to exemplar-like representations as more data are acquired, and explaining human categorization performance within both regimes [10]. Sandborn, Griffiths, and Navarro [11] also explored how rational approximation methods can be used to approximate category learning.

Inverse planning models, including Bayesian planning, have been shown to accurately predict human behavior in IDM tasks. In a series of experiments, inverse

planning models accurately captured quantitative human inferences about agents' goals, joint beliefs and desires [12]; and social relationships with other agents, such as "chasing", "fleeing", "helping", or "hindering" [13,14].

2.2 Model-Based fMRI

Machine learning applications to neural and fMRI data, including correlation-based classifiers, support vector machines (SVMs), and Gaussian Naïve Bayes (e.g., [15,16]), have been used for more than 10 years to better analyze this complex data. Neural structures and data are inherently hierarchical, and therefore a number of applications of hierarchical (a.k.a., deep) learning techniques such as hidden Markov models (HMMs), dynamical components analysis, and dynamic Bayesian networks have been applied with some success [17]. In particular, Janoos et al. [18] used HMM learning techniques to learn spatio-temporal patterns in fMRI data, deriving additional predictive strength from the representation of temporal information in the model. While these deep learning techniques have been shown to have distinct advantages over shallow learning algorithms, they have limited application due to the difficulty of tuning the algorithms, their high learning and inference time, and their resulting inability to learn more than two hidden layers.

Model-based fMRI [3,19] seeks to identify neural representations and processes by temporally correlating neural activation (i.e., fMRI voxels) with the states and outputs of computational models. For example, a feature classification IDM task may ask participants to identify to which category from a specified set an object belongs. In this case, a model-based fMRI may reveal that when categories overlap minimally in features, activation in the basal ganglia is strongest during classification. This possible result indicates that there are neural processes that related to the category overlap as computed by the model, and that this computation reflects aspects of the underlying neural structure. Similarly, the model-based analysis can be extended to cognitive and behavioral data (e.g., intuitive decisions and response times) to reveal which components of the model most predict behavior and aspects of cognition. Using the model-based analysis, model parameters can be fitted to the data and compare and select models based on their correlation with observed results.

3 Towards Model-Based fMRI with Sum-Product Networks

3.1 Sum-Product Network Overview

Recently, SPNs were developed as a new deep learning architecture [4]. Their key feature is that reasoning is linear in the size of the network. This makes learning and inference significantly faster than for existing deep architectures. In turn, this leads to the ability to learn much deeper networks, capturing more of the structure of the data.

Formally, SPNs derive from the network polynomial of graphical models [20]. The network polynomial is based on representing each variable in the network using a set of indicator functions. An indicator function indicates that a particular state of a

variable holds. Darwiche showed the probability distribution described by a graphical model can be represented by a polynomial in the indicator functions. The probability of a state of the variables can be computed by setting the corresponding indicators and computing the value of the polynomial. The probability of evidence, which is an assignment of values to a subset of variables, can be computed by setting the appropriate indicators of those variables and allowing other variables to take on all possible values by setting all indicators.

Darwiche further developed the concept of an arithmetic circuit [21], which is a compact representation of the network polynomial. A SPN is a generalization of arithmetic circuits to allow it to represent polynomials that are not necessarily derived from a graphical model. A SPN is a directed acyclic graph with three kinds of nodes. Leaf nodes represent indicator functions. Product nodes represent the product of their children. Sum nodes have weights from them to each of their children, and represent the weighted sum of their children.

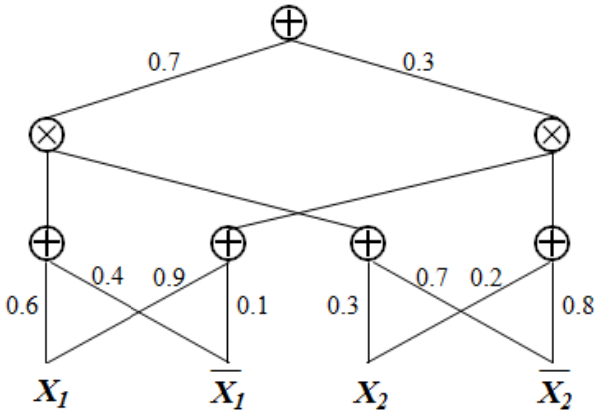


Fig. 1. Example SPN

An example SPN is shown in Fig. 1. There are two variables, 1 and 2, each of which can take on the values true and false. As a shorthand, X1 represents the indicator function for variable 1 taking the value true, while $\overline{X_1}$ represents the indicator function for variable 1 taking the value false, and similarly for variable 2. To compute the probability that variable 1 is false and 2 is true, we set the indicators $\overline{X_1}$ and X2 to 1 and the other two indicators to 0. We then propagate values from the indicators, through the sum nodes, the product nodes above them, and on to the root. To compute the probability of the evidence that variable 2 is true, we set the indicators X1, $\overline{X_1}$, and X2 to 1. We can now also compute the conditional probability that variable 1 is false given that variable 2 is true. All these computations take time linear in the size of the network. In addition, computing the most likely state of variable 1 given that variable 2 is true can also be accomplished in linear time by replacing the sum nodes with maximizations.

SPNs can be extended beyond Boolean variables to include continuous variables by using integral of the variable's probability density function instead of sum nodes at the lowest level of the network. SPN learning begins with a dense network with all possible relationships. Then, as weights are discovered to be close to 0, edges are removed to simplify the network.

3.2 Approaches to Applying Sum-Product Networks to Model-Based fMRI

This section describes some initial applications of the SPN representation and learning algorithm to fMRI and model-based fMRI data analysis.

fMRI data for a single participant consists of time-series data on brain activation for a set of voxels, which are 3 dimensional areas in the brain. The activation signal is the Blood Oxygen Level Dependent (BOLD) signal that measures the level of oxygenation of the blood in each voxel. The BOLD signal is captured across the region of interest at one or two times per second, forming a developing picture of brain activation over time. $3mm^3$ voxel regions are common, resulting in images consisting of thousands of voxels at each time slice.

To apply SPNs directly to fMRI data, each voxel or region of voxels is assigned a variable in the SPN. The variable is set to continuous value of the voxel at the current time step. In addition, variables representing the experimental conditions and behavioral measures are created. For example in a feature classification IDM task, variables may indicate the category of the current object, the observable features of the object, and the category of classification reported by the participant. The SPN is created with a dense network, and the SPN learning algorithm is applied to learn the most likely structure of the network given the data. The resulting SPN describes the learned relationship between the experimental conditions, participant behaviors, and fMRI data. It is a model of how activation instantaneously correlates with these conditions.

To apply SPNs to model-based fMRI, the model state and predictions are included in the SPN model. Using the same technique as above, the SPN learning algorithm detects the relationships between the computational model, experimental conditions, participant behaviors, and fMRI data. For example, a model-based analysis of a Bayesian network in a feature classification IDM task would include the fMRI data, category of the current object, the observable features of the object, and the category of classification reported by the participant, as well as the elements of the Bayesian model: the observed values, the intermediate nodes in the network, and the output nodes in the network. The learned SPN describes the relationship between the experimental conditions, participant behavior, fMRI data, and model predictions. It can answer questions from the learned network structure such as "What activation is most likely given a specific state of the Bayesian model?".

4 Discussion and Future Work

The application of SPNs to model-based fMRI raises several questions for the SPN approach and the nature of the fMRI data in IDM tasks. First is the issue of

computational complexity. Given the large volume of data collected by the fMRI scanner, SPNs that represent each voxel will contain thousands or tens of thousands of continuous-value variables, with a correspondingly exponential number of links. Although SPNs have been shown to be orders of magnitude faster to learn in some initial tasks, this size may prove computationally slow or intractable. Techniques for data reduction and pre-processing may be desirable to reduce the initial SPN network to feasible sizes. Approaches may include restricting analysis to regions of interest, decreasing resolution through averaging of activation, or reducing initial model size through excluding relationships from the analysis (i.e., cutting links in the initial SPN model).

Second, many intuitive tasks require the integration of information over time. For example, intuitive sequence learning (ISL) tasks directly require this ability. An analysis of only instantaneous fMRI data will be unable to correctly predict and identify the key relationships. The SPN model may be expanded to include temporal information by directly representing this information in variables (e.g., a variable that indicates the number of milliseconds since a stimulus), including simple "state" variables in the SPN to account for changes in the internal state of the system, or expanding the SPN representation to account for temporal sequences in a general. This last approach is similar to the expansion of Bayesian nets into dynamic Bayesian nets (DBNs) [22], which has lead to the application of these models in many new domains and situations.

5 Conclusions

This paper presents the first steps in a new technique to validate computational models of intuitive decision making (IDM) with functional magnetic resonance imaging (fMRI) human-performance data. Our approach builds upon previous successes in deep learning approaches to learn patterns in fMRI data and combines an improved deep learning approach, sum-product networks, with model-based fMRI analysis. We discuss several research questions that arise from this application and note the need for future work. Model-based fMRI analysis with SPNs are a promising new approach to analyzing and validating computational models of IDM.

References

1. Lieberman, M.D.: Intuition: A Social Cognitive Neuroscience Approach. Psychological Bulletin 126, 109–137 (2000)
2. Lieberman, M.D., Chang, G.Y., Chiao, J., Bookheimer, S.Y., Knowlton, B.J.: An Event-Related FMRI Study of Artificial Grammar Learning in a Balanced Chunk Strength Design. Journal of Cognitive Neuroscience 16(427), 438 (2004)
3. O'Doherty, J.P., Hampton, A., Kim, H.: Mode-Based FMRI and Its Application to Reward Learning and Decision Making. Annals of the New York Academy of Sciences 1104(1), 35–53 (2007)

4. Poon, H., Domingos, P.: Sum-Product Networks: A New Deep Architecture. In: Proc. Uncertainty in Artificial Intelligence (2011)
5. Nissen, M.J., Bullemer, P.: Attentional Requirements of Learning: Evidence From Performance Measures. Cognitive Psychology, 191–232 (1987)
6. Seidler, R.D., Purushotham, A., Kim, S.G., Ugurbil, K., Willingham, D., Ashe, J.: Neural Correlates of Encoding and Expression in Implicit Sequence Learning. Experimental Brain Research 165(1), 114–124 (2005)
7. Neal, R.M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics, 249–265 (2000)
8. Sanborn, A.N., Griffiths, T.L., Navarro, D.J.: A More Rational Model of Categorization, 726–731 (2006)
9. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
10. Griffiths, T.L., Canini, K.R., Sanborn, A.N., Navarro, D.J.: Unifying Rational Models of Categorization Via the Hierarchical Dirichlet Process, 323–328 (2007)
11. Sanborn, A.N., Griffiths, T.L., Navarro, D.J.: Rational Approximations to Rational Models: Alternative Algorithms for Category Learning. Psychological Review 117(4), 1144–1167 (2010)
12. Baker, C.L., Saxe, R.R., Tenenbaum, J.B.: Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution (2011)
13. Baker, C.L., Goodman, N.D., Tenenbaum, J.B.: Theory-Based Social Goal Inference, 1447–1452 (2008)
14. Ullman, T.D., Tenenbaum, J.B., Baker, C.L., Macindoe, O., Evans, O.R., Goodman, N.D.: Help or Hinder: Bayesian Models of Social Goal Inference. In: Neural Information Processing Systems Foundation (2009)
15. Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., et al.: Predicting Human Brain Activity Associated With the Meanings of Nouns. Science 320(5880), 1191–1195 (2008)
16. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., et al.: Learning to Decode Cognitive States From Brain Images. Machine Learning 57(1), 145–175 (2004)
17. Duan, R., Man, H., Jiang, W., Liu, W.C.: Activation Detection on Fmri Time Series Using Hidden Markov Model, 510–513 (2005)
18. Janoos, F., Machiraju, R., Singh, S., Morocz, I.: Spatio-Temporal Models of Mental Processes From FMRI. Neuroimage (2011)
19. Daw, N.D., O'Doherity, J.P., Dayan, P., Seymour, B., Dolan, R.J.: Cortical Substrates for Exploratory Decisions in Humans. Nature 441(7095), 876–879 (2006)
20. Darwiche, A.: A Differential Approach to Inference in Bayesian Networks. Journal of the AC 50(3), 280–305 (2003)
21. Darwiche, A.: Modeling and Reasoning With Bayesian Networks. Cambridge University Press, Cambridge (2009)
22. Murphy, K.: Dynamic Bayesian Networks: Representation, Inference, and Learning. Ph.D. Dissertation U.C. Berkeley (2002)

Human Memory Systems: A Framework for Understanding the Neurocognitive Foundations of Intuition

Paul J. Reber*, Mark Beeman, and Ken A. Paller

Department of Psychology, Northwestern University, Evanston, U.S.A.
{preber,mjungbee,kap}@northwestern.edu

Abstract. A neurocomputational framework is described for characterizing how intuitive and deliberate processing are accomplished in the human brain. The framework is derived from memory systems theory and supported by research findings on contrasts between implicit versus explicit (nonconscious versus conscious) memory. Implicit intuition and deliberate deduction depend on separate types of memory supported by distinct brain networks. For optimal decision making, training should be designed to accommodate the operating characteristics of both types of memory. Furthermore, reliance on explicit memory can inhibit the use of implicit intuition, so training must facilitate effective interactions between the two types of mechanism. To aid investigations of these effects, we introduce a Mixture-of-Experts model that characterizes the interaction between memory systems — the PINNACLE model (Parallel Interacting Neural Networks Competing in Learning). This model captures the separate neural networks that reflect implicit and explicit processing, as well as their interaction, and it can thus guide the development of training approaches to maximize the benefits of concurrent use of both intuition and deliberation in decision making.

Keywords: Intuition, decision making, implicit, explicit, memory systems, cognitive neuroscience, cognitive modeling.

1 Introduction

A fireman in Cleveland cleared his team from a fire scene because he "sensed" that something was odd about the situation. Indeed, the floor was about to collapse because of a raging fire below. The lieutenant fireman who saved his men was not aware of the danger in the usual sense, but rather he was observant enough and skilled enough to know that something was not right. He acted on that indication before consciously realizing what wasn't right or what danger was present. At first he thought it was ESP. Only much later did he begin to understand the clues he had sensed. [1-2].

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 474–483, 2013.

This story exemplifies the successful use of intuition in a high-pressure problem-solving environment. The profound action that saved these firefighters can be credited to implicit processing of the environmental cues, leading to escape from an imminent catastrophe. Decades of research on implicit learning have shown that our brains possess an array of mechanisms for automatically extracting information from the environment without our awareness [3]. The results of this implicit learning often appear as an intuition or a "sixth sense" about the current situation. Intuition typically emerges with no awareness of the mental events leading to it, which fits with our conjecture that implicit memory is critical in producing trustworthy intuition. Our framework builds on a substantial body of research on implicit memory in order to elucidate how this distinct yet powerful type of processing can support reliable decision-making.

Our prior research has identified neural correlates of implicit memory that we can measure to reveal implicit influences in complex tasks [4-5], and the emergence of implicit information when people solve with sudden insight [6]. We have also described a computational model to characterize the interaction of implicit and explicit processing [7]. That model, PINNACLE (Parallel Interacting Neural Networks for Competitive Learning) will be used as a basis for characterizing the neurocognitive processes involved in intuitive decision-making influenced by implicit processes. Two key features of this model are: (1) it incorporates separate processing streams for explicit deliberative processing versus implicit intuitive processing, and (2) it includes a neurocognitive architecture to test hypotheses about how these types of processing compete with each other, or conjointly produce decisions. This model makes distinct predictions about the neural basis of interactions among types of memory that can be explored and tested with functional neuroimaging approaches.

Laboratory studies of implicit learning have typically found the greatest influence of implicit knowledge when people feel they are just guessing. When implicit and explicit processing are pitted against each other in experiments, the systems often appear to compete such that only one system can influence behavior. For instance, when explicit problem solving is actively engaged, a contribution from implicit intuition is less likely, suggesting that deliberate processing can actively block the use of intuitive knowledge. Although such an arrangement seems suboptimal from a human information-processing perspective, it may reflect a characteristic of the human neural architecture that needs to be understood in order to enable the best use of implicit intuition. Findings of competition among memory and decision-making systems raise important questions about how to optimize teaching and training programs to maximize the ability of a trainee to incorporate both sources of information effectively.

The PINNACLE framework is constructed as a Mixture-of-Experts model in which independent processing streams feed information forward to a high-level cognitive process, which resolves competition and selects a response. A special feature of this model is that one stream operates outside awareness so that subjective introspection yields limited information about how this information affects behavior. Of note, the high-level decision process can function to inhibit the use of either type of information, consistent with empirical observations of competition between memory

types. We hypothesize that this meta-cognitive process can be separately trained to foster better use of both types of information and reduce inter-system competition between types of memory.

This framework enables us to test critical hypotheses about people who act based on intuition, as did the fireman in Cleveland. In his case, his prior learning about dangerous environments apparently enabled a novel pattern of cues to prime the suspicion that the floor was about the collapse. Just before this happened, what explicit processing was also engaged? How did implicit information emerge at the critical moment, and avoid suppression, to allow him to take the life-saving action? Why do others fail to be heroes in such circumstances?

If we looked into the brain of the fireman just before he saved his team, we would expect to see neural activity associated with implicit environmental pattern detection. Yet, the fireman thought that at that moment he was supernaturally able to predict the future. Given the competitive nature of implicit and explicit processing, we predict a dearth of neural activity in regions responsible for the deliberate processing of environmental cues to danger. Rather, the implicit processing of those cues likely predominated. In some domains, however, intuitive processing appears to coexist with explicit processing with less detrimental competition. During problem solving, for example, participants can be actively and explicitly searching for solutions when an insight suddenly emerges [6]. What factors facilitate the emergence of intuitive strokes of genius?

2 Mixture of Experts Model: PINNACLE

A key challenge for understanding how we use intuition in problem solving is that intuition depends materially on the result of implicit learning mechanisms that are represented in separate neural systems from deliberative problem solving. The proposed research addresses this challenge using a computational modeling approach that incorporates multiple information processing streams that are combined at the final decision process. The general PINNACLE framework is a Mixture-of-Experts (MoE) cognitive architecture, shown in Figure 1.

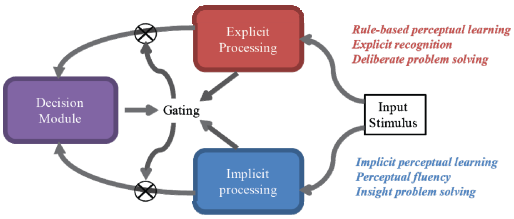


Fig. 1. General Mixture-of-Experts Cognitive Architecture of PINNACLE. Information flows from right to left from input through two parallel processing streams, explicit (upper) and implicit (lower). Three examples of explicit and implicit processes assessed in laboratory empirical studies are shown. The results from these independent processes are evaluated in a final Decision Module. Gating processes reflect competition and potential inhibition between types of processing.

Under this modeling approach, environmental information (stimulus input) is available to implicit and explicit processing streams that each operate independently in different areas of the brain. Information feeds forward to a decision module where a single behavioral response is selected as an action or decision. In addition, the model allows for a gating process to inhibit or enhance processing in one stream or the other. This architecture captures situations where strategic factors cause decision making to be locked into one mode or another—such as when a person is exclusively focused on explicit processing and no influence of implicit processing or intuition is evident. In this case, implicit processing is dormant due to inhibitory gating from the explicit process. Yet, this situation can theoretically be remedied via training to block the gating process so that implicit information can be used.

Most theories of problem solving and decision making have focused largely on processing represented in the explicit processing stream that reflects conscious, deliberative analysis of input. The effects of implicit processing appear occasionally as a sudden intuition that, when accurate, reflects the operation of nonconscious processing or memory. Decades of memory systems research have established the existence of these multiple types of processing in the human brain and provided hypotheses about the neurocognitive basis of each type of memory. However, very little research has examined the important practical questions of how information across regions may be effectively combined to guide decision making.

Three examples of how the PINNACLE framework is applied to laboratory studies of implicit and explicit processing are described here. Each example uses a different type of complex decision that can be made based on either implicit or explicit processing. Capturing these complex and interacting processes in our framework shows how the neurocognitive foundation of implicit intuition can be modeled.

2.1 Applying PINNACLE to Perceptual Learning

The PINNACLE model was first developed and applied to studies of perceptual skill learning in a visual category-learning paradigm. The visual category-learning paradigm presents participants with sine-wave gratings organized into two unknown categories that are learned during an experimental session via trial-and-error feedback. Two conditions are used to separately examine deliberate rule-based processing and implicit (termed "information-integration") category learning. Conditions conducive to RB learning are created by using a category structure that can be easily described as a rule about the stimuli. The rule is discovered by participants readily, leading to subsequent explicit rule-based category judgments. When the categorization rule requires using information across stimulus dimensions and does not lend itself to an easily verbalized rule, learning depends on implicit memory and accurate performance is not accompanied by awareness of the category structure.

To simulate both types of behavior, PINNACLE was developed with two core component processes: a rule-based learning system and an information-integration learning system. External stimuli feed information into these two parallel processing streams, which propagate information to a Decision Module, where the categorization

decision response is made [7]. Each of the processing streams (the internal "experts") is simulated using a Decision Bound Theory (DBT) mathematical model that produces a category membership estimate learned from experience, but that is constrained to only consider either rule-based or information-integration hypotheses. The DBT formalism provides an estimate of the probable category membership of a stimulus as a function of its distance in perceptual space from the category boundary, and weighted by a perceptual shaping parameter that decreases the strength of the position near the boundary conditions, where uncertainty is higher [8-9]. At the beginning of a simulated experiment, the structure of the category to be learned is not known, and both internal models attempt to learn the category via feedback. On each trial, both systems update internal representations of the category in order to improve future predictions by an error-minimizing adjustment to the current state.

The modeling process operates in two steps. In the first step, a multi-system computational model is fit to overall group behavior to establish a basic working model. In Nomura and Reber [7], we showed that groups of model simulations fit average human behavior for both kinds of category learning without needing any advance knowledge on the type of category being learned. For the second step, each individual's performance within a learning session is fit using maximum likelihood estimation to provide a model of their cognitive state during each response trial, for both the internal implicit and explicit learning processes. Free parameter values are identified that maximize the likelihood of each response in the observed sequence of behavior using a downhill simplex optimization method shown to be effective for this process [7]. We can then identify key behavioral choice moments from data collected during functional neuroimaging based on predictions of the mental state of the participant and the estimated roles of the implicit and explicit processing streams. In Figure 2, brain activity indicating the neural correlates of the separate implicit and explicit processing streams and with the process of resolving these competing sources of information is shown derived from this method.

The application of the PINNACLE framework to implicit and explicit processes in visual category learning provides a demonstration of how this modeling approach can be used to establish the neurocognitive foundations of both types of memory in complex decision making. By providing the ability to assess neural activity across both types of processing, we can observe when and how implicit intuition can be effectively brought to bear on explicit processing. In addition, when competitive interactions among types of memory reduce the use of implicit intuition, the neural basis of this effect will provide a measure of effectiveness of potential interventions to reduce competition and improve training.

2.2 Applying PINNACLE to Recognition Memory

Another example of a decision process that is potentially affected by both implicit and explicit processing is that required to make a judgment about prior occurrence (e.g., have you see this stimulus previously?). In a recognition memory test, processes of implicit and explicit memory can both contribute to accurate performance [9]. Although a recognition judgment is conventionally taken to be a straightforward test

of explicit memory, our recent work has shown that a correct response can also be produced based on a contribution of visual perceptual fluency. Explicit recognition judgments use a recognition cue (such as a word that may have been presented in a prior study list) to elicit explicit retrieval for the same item from the past (which may in some cases also include recall of relevant contextual features of a prior learning episode). However, the recognition cue can also be processed more efficiently because of the prior episode. This repetition-based efficiency is often ascribed to a boost in the fluency of perceptual processing of the cue. Responses that are seemingly guesses can actually be based on fluency signals, when an old item is selected in a recognition test without any awareness of memory for the relevant past experience.

In a series of studies [5,10-12] we have shown that we can boost the implicit memory contribution to recognition with the following set of procedures. Memory for single kaleidoscope images (each created with a unique algorithm using three colors) was tested using a two-alternative forced-choice test. The correct choice was a stimulus seen 1-2 minutes earlier; the foil choice was a very similar stimulus creating by altering the algorithm slightly, such that the decision was very difficult. Sets of stimuli were learned under divided-attention conditions, in which elaborative encoding was limited due to the concurrent demands of an auditory working-memory task. During the test, participants were encouraged to guess, and choices were made quickly using a 2-second response-signal procedure. Results were unlike standard findings for explicit memory, in that recognition accuracy was higher after divided- than full-attention encoding, and higher for guess responses compared to confident or familiarity responses. In addition, electrophysiological evidence implicated implicit perceptual fluency in accurate recognition guesses in these conditions that emphasized the use of implicit memory as opposed to explicit retrieval.

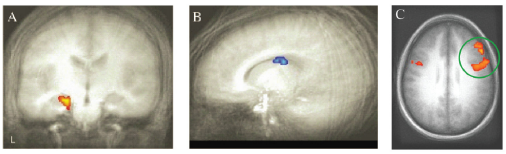


Fig. 2. Neural correlates of key brain systems involved in categorization decisions. (A) Medial temporal lobe activity associated with explicit memory for prior examples. (B) Posterior caudate activity rrelates of key brain syste associated with implicit learning. (C) Dorsolateral prefronal cortex activity associated with resolving competition between implicit and explicit processing.

On any trial in this recognition test, a correct response can be mediated by visual fluency or by explicit retrieval. We observe brain activity associated with either type of memory in EEG signals, computed by averaging trials for different judgments together. One indication of the type of response comes from metamemory judgments; participants can either indicate confidence in their response (i.e., conscious, explicit retrieval) or they can indicate a response made on no known basis whatsoever (guess). These highly accurate guess responses are what we term “implicit recognition” [9]. Judgments may also be made on the basis of implicit fluency signals in a variety of other decision-making circumstances.

By creating conditions wherein implicit recognition occurs on a large proportion of trials, the PINNACLE framework provides a method for examining the neurocognitive foundations of both types of processing and also potential interactions between the two types of memory. A key question is whether and how explicit retrieval blocks or interferes with the use of implicit knowledge. A focus on explicit memory retrieval appears to limit the extent to which implicit information is available for making a memory decision; both the number of guesses and the accuracy of those guesses is reduced by changing the instructions to emphasize confident responding [11] or by interfering with brain activity in prefrontal cortex [Lee, Blumenfeld, & D'Esposito, unpublished manuscript]. Paradigms that overcome this inhibitory (gating) effect will serve as a model for training the ability to simultaneously use both implicit and explicit memory in complex decision-making.

2.3 Applying PINNACLE to Insight Problem Solving

The third example domain for examining interactions between implicit and explicit processing is the laboratory study if insight-driven problem solving. In general problem solving, people can achieve solution using analytic processing, sudden creative insight, or both [6,13]. Analytic solving relies heavily on step-by-step processing and deliberate manipulation of consciously accessible information with explicit awareness of the contents and strategies engaged. In contrast, insight solving occurs when a person suddenly becomes aware of a solution, without conscious access to the solving process. Thus, compared to analytic solving, insight is more influenced by implicit memory and implicit processes generally.

Recently we've examined and manipulated factors that modulate the degree to which analytic and insight processes contribute to solving problems. In order to elicit robust numbers of both analytic and insight solutions, we've most often presented people with a large number of Compound Remote Associate (CRA) problems, in which they view three problem words (e.g., pine, crab, sauce), and must produce a solution word that can form familiar compounds or two-word phrases with each of the problem words (apple: pineapple, crabapple, apple sauce) [14]. On average, people can solve about half of these problems, and about half of the solutions occur with each type of solving. In numerous studies, participants indicate how they solved each problem, by analysis or insight. Different solution types are associated with changes in behavior, neural activity, blinks, and eye movements, all indicating that the participants engaged in different processes prior to solution. Indeed, insight and analytic solving are associated with different forms of attention prior to engaging each problem [15], and even different baseline brain activity [16].

Moreover, mood differentially affects insight and analytic solving, with positive mood facilitating insight, most likely via changes in anterior cingulate cortex that modulate cognitive control [17]; and separate visual tasks that encourage highly focused external attention facilitate analytic solving, whereas visual tasks that encourage internal attention facilitate insight solving [18]. Using the PINNACLE framework, we can characterize these effects as emphasizing processing within either the explicit, deliberative processing stream or the implicit, intuitive processing that

leads to sudden insight. Emphasis on one type of problem solving approach may be reflected as directly increasing neural activity within one of the processing streams or may be reflected in high-level decision making processes that indicate a strategic decision to rely on step-wise problem solving or to anticipate a sudden flash of insight. By examining the neurocognitive foundations of these interacting processes, the problem solving paradigm provides a useful model of the roles of implicit and explicit memory in a cognitively complex domain.

3 Designing Interventions to Improve Use of Intuition

The key questions for improving the use of intuition are focused on the gating and decision-making mechanisms that are engaged during integration of information between the implicit and explicit processing streams. A variety of approaches aimed at increasing reliance on implicit intuition are derived from our prior research on implicit learning. To evaluate these approaches, we can quantify the improvement in performance using these paradigms. In addition, the PINNACLE modeling approach makes testable predictions about how the underlying neural activity patterns are changed by successful training interventions.

For instance, to target improvements in the operation of gating and reducing interfering competition, we could attempt to improve intuitive decisions using metacognitive strategies that avoid overshadowing of implicit information by explicit processing. That is, we can reduce dependence on highly focused external attention. To boost the impact of implicit processing, we can train participants to induce inward-looking attention to quiet internal activations and associations [18]. To do so, we can combine methods for inducing inward attention (e.g., voluntary eye-blinks and overt eye fixations away from problem stimuli) with feedback based on both successful implementation of the attention strategy and successful intuitive decisions.

Another approach is to use trial-by-trial feedback in order to give participants a greater ability to internally monitor their experience of implicit visual fluency signals in recognition judgments, using reinforcement-learning mechanisms. This approach is based on the idea that trainees can gradually learn to use subtle visual fluency cues more often, such that implicit intuition plays a greater role in complex decision making or problem solving. The feasibility of this method to train participants to use fluency this way is supported by recent findings from exposing subjects to a situation in which previously unstudied items were less visually fluent than studied items—and reinforcing this connection with trial-by-trial feedback [19-20]. Whereas familiarity is typically attributed to old items because they are, on average, more fluently processed than new items, this manipulation led to a temporary reversal such that subjects acquired a tendency to attribute familiarity to items with less fluency. By analogy, trainees should be able to learn the contingencies between the beneficial use of visual fluency and positive feedback for correct decisions—and these habits will generalize to other circumstances wherein implicit processing can be beneficial.

A third approach to improving the use of implicit intuition is based on the hypothesis that people can be trained to more strongly weight the implicit processing

stream during decision making. Such training would encourage the use of implicit knowledge. This hypothesis suggests that the use of implicit intuitive knowledge could be enhanced in scenario-based training based on rapid decision making with ambiguous cues by providing pre-training with tasks that rely on implicit learning. Experience with successful implicit learning would then be used as a training enhancement to increase the ability to integrate knowledge across information processing systems, producing increased decision-making ability.

These three ideas reflect examples of how it is possible to use information about the neurocognitive foundations of implicit intuition in decision making in order to learn how to better use intuition. As we better understand the neural processes associated with memory systems in complex decision making, it is likely that a wide range of additional ideas for training interventions can be developed.

4 Summary and Conclusions

Our computational framework, PINNACLE, provides a neurocognitive foundation for studies examining the interacting roles of intuition and planned, deliberate processing in complex decision-making environments. By connecting implicit and explicit processing directly to neural circuitry, we can develop strategies for studying these processes individually and also tackle the challenge of how these two types of memory interact. Training effects can therefore be attributed to behavioral change reflecting one type of memory or the other. Experts with strong intuitions based on implicit learning from extensive experience rely on a different type of neural processing than do individuals who have learned an explicit rule. In addition to simulation-based training to provide an analog to situational experience, enhancing the ability to apply this intuition alongside explicit rules will also be necessary to bring trained intuition to bear on complex real-world problems.

References

1. Gladwell, M.A.: Blink: The power of thinking without thinking, p. 122. Little, Brown & Co., Boston (2005)
2. Van Hecke, M., Callahan, L., Kolar, B., Paller, K.A.: The Brain Advantage, p. 201. Prometheus Books, Amherst (2009)
3. Reber, P.J.: Cognitive neuroscience of declarative and nondeclarative memory. In: Benjamin, A.S., De Belle, J.S., Etnyre, B., Polk, T.A. (eds.) Advances in Psychology, vol. 139, pp. 113–123. North-Holland (2008)
4. Nomura, E.M., Maddox, W.T., Filoteo, J.V., Ing, A.D., Gitelman, D.R., Parrish, T.B., Mesulam, M.M., Reber, P.J.: Neural correlates of rule-based and information-integration visual category learning. Cerebral Cortex 17, 37–43 (2007)
5. Voss, J.L., Paller, K.A.: An electrophysiological signature of unconscious recognition memory. Nature Neuroscience 12, 349–355 (2009)
6. Jung-Beeman, M., Bowden, E.M., Haberman, J., Frymiare, J.L., Arambel-Liu, S., Greenblatt, R., Reber, P.J., Kounios, J.: Neural activity observed in people solving verbal problems with insight. Public Library of Science – Biology 2, 500–510 (2004)

7. Nomura, E.M., Reber, P.J.: Combining computational modeling and neuroimaging to examine multiple category learning systems in the brain. Brain Sciences 2, 176–202 (2012)
8. Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., Waldron, E.M.: A neuropsychological theory of multiple systems in category learning. Psychological Review 105, 442–481 (1988)
9. Voss, J.L., Lucas, H.D., Paller, K.A.: More than a feeling: Pervasive influences of memory processing without awareness of retrieval. Cognitive Neuroscience 3, 193–207 (2012)
10. Voss, J.L., Baym, C.L., Paller, K.A.: Accurate forced-choice recognition without awareness of memory retrieval. Learning & Memory 15, 454–459 (2008)
11. Voss, J.L., Paller, K.A.: What makes recognition without awareness appear to be elusive? Strategic factors that influence the accuracy of guesses. Learning & Memory 17, 460–468 (2010)
12. Vargas, I.M., Voss, J.L., Paller, K.A.: Recognition based on lateralized perceptual fluency. Brain Sciences 2, 22–32 (2012)
13. Kounios, J., Beeman, M.: The Aha! moment: The cognitive neuroscience of insight. Current Directions in Psychological Science 18, 210–216 (2009)
14. Bowden, E.M., Jung-Beeman, M.: One hundred forty-four Compound Remote Associate Problems: Short insight-like problems with one-word solutions. Behavioral Research, Methods, Instruments, and Computers 35, 634–639 (2003)
15. Kounios, J., Frymiare, J.L., Bowden, E.M., Fleck, J.I., Subramaniam, K., Parrish, T.B., Jung-Beeman, M.: The prepared mind: Neural activity prior to problem presentation predicts solution by sudden insight. Psychological Science 17, 882–890 (2006)
16. Kounios, J., Fleck, J., Green, D.L., Payne, L., Stevenson, J.L., Bowden, E.M., Jung-Beeman, M.: The origins of insight in resting-state brain activity. Neuropsychologia 46, 281–291 (2008)
17. Subramaniam, K., Kounios, J., Parrish, T.B., Jung-Beeman, M.: A brain mechanism for facilitation of insight by positive affect. Journal of Cognitive Neuroscience 21, 415–432 (2009)
18. Wegbreit, E., Suzuki, S., Grabowecky, M., Kounios, J., Beeman, M.: Visual attention modulates insight versus analytic solving of verbal problems. Journal of Problem Solving 4(2), Article 5 (2012)
19. Unkelbach, C.: The learned interpretation of cognitive fluency. Psychological Science 17, 339–345 (2006)
20. Olds, J.M., Westerman, D.L.: Can fluency be interpreted as novelty? Retraining the interpretation of fluency in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 38, 653–664 (2012)

Modeling Cues for Intuitive Sensemaking Simulations

Sae Schatz and Kathleen Bartlett

MESH Solutions, LLC (a DSCI Solution), Orlando, Florida, USA
{sschatz,kbartlett}@mesh.dsci.com

Abstract. Modern military personnel must not only possess typical warfighting abilities; they must also be able to rapidly perceive, understand, and then respond to a range of ambiguous behavioral, social, and cultural stimuli. In other words, personnel must have sociocultural sensemaking skills—preferably *intuitive* sensemaking skills that allow them to act with the utmost agility. This paper begins by discussing sensemaking, sociocultural pattern recognition, and expertise-based intuition. It briefly describes training approaches for these constructs, as well as training for the integrated concept. Instructional simulations could facilitate such training. However, for simulations to effectively support this subject matter, they must be able to replicate realistic patterns of life, from the subtle characteristics of human body language to the emergent behaviors of crowds. That is, they must provide accurate, nuanced cues to which the trainees can react. This paper closes by discussing our ongoing work to address this gap by modeling realistic cues in a simulation.

Keywords: sensemaking, intuition, patterns of life, simulation, military training, cognitive readiness.

1 Introduction

Since 2001, the United States has engaged in an unconventional military conflict defined by a range of counterterrorism, counterinsurgency, peacekeeping, and infrastructure-building initiatives. Consequently, modern military personnel must not only possess typical warfighting abilities, but they must also be able to rapidly perceive, understand, and then respond to a range of ambiguous behavioral, social, and cultural stimuli. In other words, personnel must develop enhanced sociocultural sensemaking skills—preferably intuitive sensemaking skills that allow them to act with the utmost agility.

Presently, our team is investigating novel approaches for training intuition, sociocultural perception, and sensemaking skills to US Marine Corps personnel. Like the US Army as well as other governmental and nongovernmental organizations, the Marines must be able to excel in potentially hostile, typically uncertain, cross-cultural settings. They must be able to enter a new location, develop a sense of its overall patterns of life, and then rapidly identify detrimental anomalies in those patterns, such as the activities of criminal networks or suicide bombers.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 484–491, 2013.

To meet these objectives, the US military has invested in various cultural training approaches; however, typical culture training lacks elements that military personnel uniquely require, such as learning how to distinguish friend from foe. Also, typical culture training does not necessarily foster sensemaking skills or anomaly detection abilities. To further complicate matters, the military must contend with highly demanding training schedules that may only allot a few weeks to learn such knowledge, skills, and attitudes.

This paper describes our current efforts to address this military training need through advanced simulation. The paper begins by defining the training objectives at a high level. Next, it describes the integrated concept of intuitive sociocultural sensemaking and our initial thoughts on how to foster its development. Finally, the paper discusses ongoing research on modeling patterns of life, which are complex sociocultural cues presented by an instructional simulation. The goal of this research is to develop sophisticated patterns-of-life computational algorithms that are capable of successfully simulating humans' individual and social behaviors. This, in turn, will support simulation-based training and practice of intuitive sociocultural sensemaking.

2 Training and Education Objectives

2.1 Intuition

Intuition is the unconscious awareness, valuation, and integration of important cues. Stated more formally, "intuition is a rapid, non-conscious cue to the existence of meaningful information detected through one or more sensory modalities" [1]. It is "a process of thinking. The input to this process is mostly provided by knowledge stored in long-term memory that has been primarily acquired via associative learning. The input is processed automatically and without conscious awareness. The output of the process is a feeling that can serve as a basis for judgments and decisions" [2]. Intuitions "arise through rapid, non-conscious, and holistic associations" [3] and involve a subjective perception of pattern, meaning, or structure [4]. In other words, intuition is the ability to put together cues at a subconscious, nonverbal level, and recognize a pattern worthy of notice before that pattern can be deliberately perceived. The feeling of intuition is the experience of knowing, without immediately knowing the reasons why [5].

Despite the subconscious facets of intuition, training and education can enhance individuals' intuitive capacities [6]. Classically, experts build their intuitive skills in particular domains through experience and implicit learning [7, 8, 9]. They learn to regulate their intuitive feelings by actively seeking feedback [10], and they selectively attending to intuitive thoughts based upon the characteristics of the problem space [6]. Therefore, intuition can be fostered by first acquiring domain experience and then developing intuition-related skills through intense deliberate practice, critical self-appraisal, and candid feedback [11]. Instructors can also help engender related skills

(e.g., divergent thinking) and factors (e.g., positive attitudes towards intuition) that enhance the likelihood of effective intuitive processing.

2.2 Sociocultural Sensemaking

Sensemaking is the ongoing process of giving meaning to one's experiences, of "structuring the unknown" [12] Stated more formally, sensemaking is the "motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively" [13]. It is the process of "placing stimuli into some kind of framework" in order "to comprehend, understand, explain, attribute, extrapolate, and predict" their individual and collective, emergent behaviors [14, p. 51]. In social settings, sensemaking supports outcomes such as sociocultural situation assessment, anomaly detection, and anticipatory thinking. It also helps individuals establish a "sense of coherence and sociocultural brokerage" [15, p.104; see also, e.g., 16, 17].

In general, sensemaking skills can be fostered through a range of instructional interventions, "[c]ombining theory, role models, action learning, feedback, and class assignments" [18, p. 13]. *Sociocultural* sensemaking is built through these general methods along with additional culture education, training, and mentorship programs [e.g., 18]. For example, the United States Marine Corps (USMC) Combat Hunter program currently teaches personnel to conduct sustained observation of social patterns, and it fosters personnel's social, cultural, and behavioral perceptual skills [20, 21]. However, for the military, the efficiency of training is an important consideration, and the Combat Hunter program is time- and personnel-intensive [20]. Therefore, automated tools must be designed that help cultivate (and ideally, accelerate) personnel's acquisition of sensemaking skills.

2.3 Intuitive Sociocultural Sensemaking

Intuitive sensemaking can be considered a conscious process, informed by subconscious intuitive mechanisms and moderated by deliberate metacognitive effort, with the intention of understanding connections, interpreting meaning, and anticipating trajectories, which support later decision making and possible actions. Further, *intuitive sociocultural sensemaking* refers to these activities as applied to human social, cultural, and behavioral stimuli. Intuition and psychosocial skills complement one another. Intuitive processing can enhance the discernment and interpretation of subtle sociocultural cues and patterns, and exploring intuition in this context addresses the timely need to enhance personnel's sociocultural abilities.

Our current work seeks to use existing and emerging technologies to support the simulation-based training of intuition and sociocultural sensemaking. This training involves four high-level sensemaking learning outcomes, three of which have immediate relevancy, here (see Table 1). For more details on these learning outcomes, see [22].

Table 1. High-level learning outcomes relevant to sociocultural sensemaking

Sociocultural sensemaking
• Taking someone else's perspective • Looking for prototypes to guide rapid recognition • Generating explanatory storylines, tying info together • Not settling for unexplained events or evidence but looking for antecedents to a situation • Mentally simulating alternative actions or outcomes • Anticipating what will happen next • Detecting an unfolding event by identifying a piece of it and inferring the rest
Developing mental baselines
• Using optics to help construct a baseline or profile • Establishing a baseline of an area to extract normalcy • Constructing a behavior profile of a person or event • Effectively and efficiently identifying leaders • Efficiently identifying anchor points and habitual areas • Constructing a behavior profile of a person or event • Orienting observation toward potentially hostile players and ignoring neutrals
Identifying anomalies
• Looking for anomalies outside of the baseline • Looking for signature behaviors via a cluster of cues • Looking for signature locations via a cluster of cues • Using appropriate criteria to make timely but accurate decisions about anomalies

3 Virtual Observation Platform

The Virtual Observation Platform (Virtual OP) is an immersive adaptive simulation-based training system, designed to instruct perceptual–cognitive skills, such as sociocultural sensemaking. In the Virtual OP, trainees observe a virtual location, such as a small town, from distal location (300–1000 meters away). The small town is represented in a virtual environment, specifically Virtual Battlespace 2 (VBS2), and trainees observe the everyday patterns of human behavior within the small town. See Figure 1.

As trainees observe the town, they learn to make sense of the patterns of activity, to establish a mental "baseline" of normal activities, identify anomalies, and, ultimately, to predict deleterious events before they occur (i.e., "left of bang"). For example, an event might involve the delivery of bomb-making supplies. The trainees would observe the terrorist cell leader meeting a pickup truck and bags of fertilizer being moved into a home. This represents an obvious anomalous cue, because these chemicals would not normally be stored in a residence.

The architecture for the Virtual OP includes a control agent system that can monitor progress within scenarios, estimate trainees' proficiency as scenarios evolve,

and invoke tailoring strategies [22]. In other words, the Virtual OP monitors student behavior during simulation-based practice and attempts to scaffold, to challenge, or to engage trainees based on the learning context. Like dedicated human tutors, these adaptive instructional technologies tailor learning content, delivery, and/or context to the unique needs of the learners. For instance, if novice trainees miss observing an anomaly, the system can scaffold their training by triggering an event (e.g., squawking chickens) to draw their attention to the delivery. For advanced trainees, the same kind of cuing event occurring in a distant area can provide a distraction. In this way, manipulating intrinsic cue quantity or cue misinformation can scaffold or challenge a perceptual skill [22].

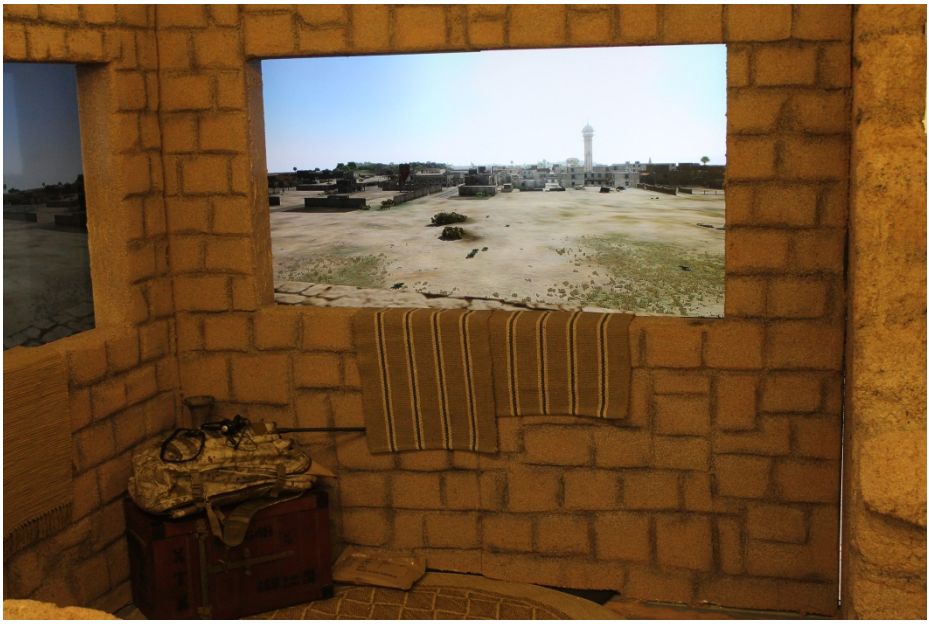


Fig. 1. Photo of part of the Virtual OP simulator

3.1 Patterns of Life

Patterns of life have been defined as the archetypal emergent properties of a complex sociocultural system [23]. These patterns originate from human behavioral and social universals. For example, universal human emotions include fear, sadness, and frustration, and universal practices include cooking food, sleeping in individual or group quarters, joining mates via rituals, identifying interactions based on kinship, exchanging greetings, wearing clothing or wraps, dividing labor, organizing relations hierarchically, making music, creating nonlinguistic symbols, and participating in death rites [24].

Archetypal patterns may involve multiple levels of observable physical interactions, communications, and routines that occur among members of social

groups, and training can involve recognition of these cultural patterns from a third-person perspective (e.g., watching interactions between genders, observing shopping behaviors). Identification of these patterns of cultural behaviors will allow an observer to monitor societal norms to establish baselines of day-to-day activities, ultimately enabling a user to detect anomalies from baselines and develop sensemaking analytical skills [21, 22, 23].

3.2 Patterns of Life Simulation

In order to effectively support sociocultural sensemaking training in the Virtual OP, cultural cues need to be developed and displayed at a depth beyond surface level. Rather than presenting an image representing a stereotypical version of a software designer's interpretation of culturally relevant details, social patterns must be researched and carefully incorporated. That is, the simulation platform must produce accurate, highly nuanced representations of patterns of life. To achieve this, accurate and scalable pattern-of-life models must be created that computationally define the patterns that trainees will attempt to make sense of. The Virtual OP must be able to replicate realistic patterns of life, from the subtle characteristics of human body language to the emergent behaviors of crowds, and these entities' individual and emergent behaviors must reinforce the training objectives and scenario narrative.

Recently developed platforms for these types of systems, such as DI-GUY, attempt to depict realistic patterns of life in immersive environments [25]. Such programs rely on a technical approach of non-linear, hierarchal software programs and use a combination of modules, such as AI Minds, SmartObject, and SmartBuilding frameworks. However, we argue that current methods for representing patterns of life fail to comprehensively address the issue. Instead, most efforts focus narrowly, at small "bubbles of life," such as the behaviors of residents on a single farm or the flow of communications within an insurgent network. These small-scale, top-down approaches have some benefits, but they are too limited to support the types of simulation required to compressively train intuitive sociocultural sensemaking.

The alternative approach of using bottom-up, agent-driven systems could support the depth of training we require; however, such systems do not provide sufficient insights into the emergent behaviors nor do they readily allow for pedagogically guided alternations. Consequently, a novel approach to modeling human individual, social, and cultural cues must be developed that can then support intuitive sensemaking simulation-based training. We are exploring one such approach, which combines bottom-up agent-based simulation with top-down supervisory control (see [23] for an overview).

By participating in simulation-based training involving patterns of life, we hypothesize that military personnel will develop enhanced—potentially intuitive--sociocultural sensemaking skills. These skills should better prepare personnel to operate in complex social contexts by teaching them to perceive sociocultural patterns, identify normal characteristics of the patterns within their areas of operation, and recognize anomalous patterns when they appear.

Within the Virtual OP simulator, we are developing and testing more flexible, scalable, and controllable approaches for generating patterns of life. Our current approach toward realizing these patterns integrates bottom-up agent-based modeling with top-down supervisory control, in order to create a dynamically manipulatable, deterministically chaotic system. Over the next three years, we will continue to refine our theory of patterns of life, as well as the corresponding implementation of that theory in our simulation system. If we are able to accurate model patterns of life in the Virtual Op, then we will be better able to support military sociocultural sensemaking, anomaly detection, and cross-cultural perceptual skills, as well as their associated intuitive processes.

Acknowledgements. This work was supported, in part, by the Office of Naval Research project N00014-11-C-0193, Perceptual Training Systems and Tools (PercepTS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

1. Warfighter Intuition workshop, Orlando, FL, February 4-5 (2009)
2. Betsch, T.: The nature of intuition and its neglect in research on judgment and decision making. In: Plessner, H., Betsch, C., Betsch, T. (eds.) Intuition in Judgment and Decision Making, pp. 3–22. Lawrence Erlbaum Associates, New York (2008)
3. Dane, E., Pratt, M.G.: Exploring Intuition and Its Role in Managerial Decision Making. Academy of Management Review 32(1), 33–54 (2007)
4. Luu, P., et al.: The neural dynamics and temporal course of intuitive decisions (n.d.)
5. Claxton, G.: The Anatomy of Intuition. In: Atkinson, T., Claxton, G. (eds.) The Intuitive Practitioner, pp. 32–52. Open University Press, Buckingham (2000)
6. Salas, E., Rosen, M.A., DiazGranados, D.: Expertise-Based Intuition and Decision Making in Organizations. Journal of Management 1-31, 941–973 (2009)
7. Agor, W.H. (ed.): Intuition in Organizations: Leading and Managing Productively. Sage Publications, Newbury Park (1989)
8. Harper, S.C.: Intuition: What separates executives from managers. In: Agor, W.H. (ed.) Intuition in Organizations, pp. 111–124. Sage Publications, Newbury Park (1989)
9. Klein, G.: Sources of power: How people make decisions. MIT Press, Cambridge (1998)
10. Hogarth, R.M.: Educating intuition. University of Chicago Press, Chicago (2001)
11. Hodgkinson, G.P.: Intuition in Organizational Decision making. Keynote address to the IST Workshop, University of Central Florida, February 4-5 (2009)
12. Colville, I.D., Waterman, R.H., Weick, K.E.: Organizing and the search for excellence: Making sense of the times in theory and practice. Organization 6(1), 129–148 (1999)
13. Klein, G., Moon, B., Hoffman, R.F.: Making sense of sensemaking I: alternative perspectives. IEEE Intelligent Systems 21(4), 70–73 (2006)

14. Starbuck, W.H., Milliken, F.J.: Executives' perceptual filters: What they notice and how they make sense. The Executive Effect: Concepts and Methods for Studying Top Managers 35, 65 (1988)
15. Glanz, L., Williams, R., Hoeksema, L.: Sensemaking in expatriation—A theoretical basis. Thunderbird International Business Review 43(1), 101–120 (2001)
16. Osland, J.S., Bird, A.: Beyond sophisticated stereotyping: Cultural sensemaking in context. The Academy of Management Executive 14(1), 65–77 (2000)
17. Vaara, E.: Constructions of cultural differences in post-merger change processes: A sensemaking perspective on Finnish-Swedish cases. Management 3(3), 81–110 (2000)
18. Ancona, D.: Sensemaking: Framing and acting in the unknown. In: Snook, S., Nohria, N., Khurana, R. (eds.) The Handbook for Teaching Leadership: Knowing, Doing, and Being, pp. 3–21. Sage publications, Thousand Oakes (2012)
19. Harvey, M., Buckley, M.R., Novicevic, M.M., Wiese, D.: Mentoring dual-career expatriates: A sense-making and sense-giving social support process. International Journal of Human Resource Management 10(5), 808–827 (1999)
20. Schatz, S., Reitz, E.A., Nicholson, D., Fautua, D.: Expanding Combat Hunter: The science and metrics of Border Hunter. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC). National Training Systems Association, Washington, DC (2010)
21. Schatz, S., Nicholson, D.: Perceptual training for cross cultural decision making (session overview). In: Nicholson, D.M., Schmorrow, D.D. (eds.) Advances in Design for Cross-Cultural Activities Part, ch. 1, pp. 3–12. CRC Press, San Francisco (2012)
22. Schatz, S., Wray, R., Folsom-Kovarik, J., Nicholson, D.: Adaptive Perceptual Training in a Virtual Environment. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56(1), pp. 2472–2476. Sage Publications (2012)
23. Schatz, S., Folsom-Kovarik, J.T., Bartlett, K., Wray, R., Solina, D.: Archetypal Patterns of Life for Military Training Simulations. In: Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL (2012)
24. Schatzki, T.R.: Human universals and understanding a different socioculture. Human Studies 26(1), 11–20 (2003)
25. Blank, B., Broadbent, A., Crane, A., Pasternak, G., Engineers, D.G.S.: Defeating the authoring bottleneck: Techniques for quickly and efficiently populating simulated environments. In: Proceedings of the 2009 IMAGE Conference (2009)

Evaluating Classifiers for Emotion Recognition Using EEG

Ahmad Tauseef Sohaib, Shahnawaz Qureshi, Johan Hagelbäck, Olle Hilborn, and Petar Jerčić*

Blekinge Institute of Technology, Karlskrona, Sweden
johan.hagelback@bth.se

Abstract. There are several ways of recording psychophysiology data from humans, for example Galvanic Skin Response (GSR), Electromyography (EMG), Electrocardiogram (ECG) and Electroencephalography (EEG). In this paper we focus on emotion detection using EEG. Various machine learning techniques can be used on the recorded EEG data to classify emotional states. K-Nearest Neighbor (KNN), Bayesian Network (BN), Artificial Neural Network (ANN) and Support Vector Machine (SVM) are some machine learning techniques that previously have been used to classify EEG data in various experiments. Five different machine learning techniques were evaluated in this paper, classifying EEG data associated with specific affective/emotional states. The emotions were elicited in the subjects using pictures from the International Affective Picture System (IAPS) database. The raw EEG data were processed to remove artifacts and a number of features were selected as input to the classifiers. The results showed that it is difficult to train a classifier to be accurate over large datasets (15 subjects) but KNN and SVM with the proposed features were reasonably accurate over smaller datasets (5 subjects) identifying the emotional states with an accuracy up to 77.78%.

1 Introduction

Humans interacting with computer applications are a part of everyday life. Similarly, emotions are a vital and constantly present part in everyday life of humans and can provide many possibilities in enhancing the interaction with computers e.g. affective interaction for disabled or people in stressful environments. As technology and the understanding of emotions are advancing, there are growing opportunities for automatic emotion recognition systems. There is much successful research on emotion recognition using text, speech, facial expressions or gestures as stimuli[1]. In this paper we focus on recognition of emotions from Electroencephalogram (EEG) signals, since this technique have the benefit of being more passive and less intrusive for the human than facial expressions or vocal intonation. The need and importance of the automatic emotion recognition from EEG signals has grown with increasing role of brain computer interface applications and development of new forms of human-centric and human-driven interaction with digital media. The asymmetry among left and right brain hemispheres are the major areas where the emotion signals can be captured[2]. According to

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 492–501, 2013.

a model developed by Davidson et al., the two core dimensions -arousal and valence- are related to asymmetric behavior of emotions. A judgment about a state as positive or negative lies under valence whereas the level of excitation (calmness, excitement) lies under arousal[3].

Human machine interaction on the base of physiological signals has been greatly investigated by previous and recent research. Of particular interest are systems that can make interpretations about psychological states based upon physiological data. Linear classifiers[4,5,6] are considered to be the most appropriate classification technology due to their simplicity, speed and interpretability. However, non-linear classifiers are considered to be the most appropriate when it comes to signal features and cognitive state[7,8].

Sequential Floating Forward Search and Fisher Projection methods are used by Picard et al. to classify eight basic emotions with 81% accuracy[9]. Lisetti and Noasoz used Marquardt Back Propagation, Discriminant Function Analysis and K-Nearest Neighbor to distinguish between six emotions and acquired classification accuracy between 71% and 83%[10]. Conati argued that probabilistic models can be developed using a methodology provided which uses various body expressions of the user, personality of user and context of the interaction[11]. Mental workload has been evaluated using Artificial Neural Networks providing mean classification accuracies of 85%, 82% and 86% for the baseline, low task difficulty and high task difficulty states respectively[12]. Fisher developed an emotion-recognizer based on Support Vector Machines which provided accuracies of 78.4% and 61.8%, 41.7% for recognition of three, four and five emotion categories respectively[4]. According to Rani et al., if the same physiological data is used then Support Vector Machines with a classification accuracy of 85.81% perform the best, closely followed by the Regression Tree at 83.5%, K-Nearest Neighbor at 75.16% and Bayesian Network at 74.03%. Performance of K-Nearest Neighbor and Bayesian Network algorithms can be improved using informative features. Support Vector Machine shows 33.3% and 25% accuracy for three and four emotion categories respectively when it comes to physiological signal databases acquired from ten to hundreds of users[13]. For more research on emotions and EEG see for example [14,15,16,17,18,19].

It is difficult to compare the results between different studies due to different experiment environments, preprocessing techniques, feature selection etc.. However, studies have shown that various factors such as preprocessing and classification techniques can strongly affect the results in terms of accuracy. Even if several methods have successfully been used to develop affect recognizers from physiological indices, it is still important to select an appropriate method in each study for the classification of EEG data to attain uniformity in various aspects of emotion selection, data collection, data processing, feature extraction, base lining, and data formatting procedures.

Several machine learning techniques have been used for classifying EEG data. Some common ones that previously have been used for EEG data associated with affective/emotional states are K-Nearest Neighbor (KNN), Regression Tree (RT), Bayesian Network (BNT), Support Vector Machine (SVM) and Artificial Neural Network (ANN).

According to an extensive survey carried out by Rani et al. KNN is one of the most widely used techniques for classifying EEG data associated with specific

affective/emotional states[13]. Yu et al. found that KNN was the most effective classifier in classifying emotion sickness from EEG data[20]. Parvin et al. claims that KNN's ability of dealing with discriminant analysis of difficult probability densities makes it very effective for classifying EEG data[21]. According to Downey and Russell, RT is largely used in medical fields to, for example, classify EEG data[22]. Brown et al. also mentions the wide use of RT for classifying EEG data[23]. BN was used with success by Matas et al. for classifying varying emotional states[24]. In their survey, Rani et al. strongly supports SVM and recommend it for accurately classifying EEG data[13]. This claim is also supported by Chen and Hou[25]. According to experiment results by Yu et al. and Huang et al., SVM provides effective and promising results for classifying EEG data[20,26]. In a study by Tangermann et al. the authors claim that SVM can show a high level of agreement on EEG data classification[27]. In a study by Ho and Sasaki ANN could accurately classify EEG data and they claim it is especially useful when a small number of electrodes are used[28]. Chen and Hou claims that ANN is an effective technique to classify EEG data due to its ability to handle noisy data efficiently[25].

These five techniques were found to be used in most of the empirical studies we have found and were considered to be suitable for the classification of EEG data associated with specific affective/emotional states based on the achieved classification accuracy. KNN and SVM seemed to be the most common ones among the classifiers with the highest attained accuracy where our interest was to achieve high accuracy over large datasets/participants.

2 Experiments

The goal of the experiments was to classify the various emotional states in subjects as they look on different pictures that are inducing strong emotions. The International Affective Picture System (IAPS) was used for this purpose. IAPS is a general picture database especially designed for experiments in emotions with normative values for valence, arousal, and dominance[29]. In these experiments we used the 2-dimensional emotional model with valence and arousal.

A total of 20 subjects (15 men and 5 women) participated in the experiment. All subjects were students of Blekinge Institute of Technology, Sweden, and aged from 21 to 35 years. The subjects were from different cultural background, nationalities and field of studies.

The EEG signals were captured from left and right frontal, central, anterior temporal and parietal regions (F3, F4, C3, C4, T3, T4, P3, P4 positions according to the 10-20 system and referenced to Cz)[30]. Based on these findings, the experiment was executed as described by Davidson et al.[3] and AlZoubi et al.[31]:

- An appropriate interface was applied for the automated projection of the IAPS emotion-related pictures.
- To compensate opening/closing of eyes 30 seconds gap was maintained before starting the experiments.
- 30 IAPS pictures (6 pictures for each emotion cluster as neutral, positive arousing/calm, negative arousing/calm) were displayed randomly for the duration of 5 seconds with a gap of a black screen between 5-12 seconds. The purpose of the

black screen duration was to reset the emotional state of subjects offering them the time to relax having no emotional content. A cross shape projection was displayed for 3 seconds before each picture to attract the attention of the subject. This process was repeated for each picture.
- A subject may feel an emotion which differs from the one expected. Therefore each subject was asked to rate his/her emotion on a Self-Assessment Manikin (SAM)[29]. Each subject rated their level of emotion on a 2D arousal and valence scale.
- Two recording sessions for 25 to 35 trials having 5 pictures, displaying each picture for 2.5 seconds were completed.
- During the whole process, subjects were directed to stay quiet and still (to realize and observe the emotion instead of mimic the facial expression) with as few eye blinks as possible to get rid of other artifacts (e.g., facial muscles).
- Fp1, Fp2, C3, C4, F3, and F4 positions were used to attain the EEG signals according to 10-20 system and all of the electrodes were referenced to Cz.

During the experiments, EEG data for each subject was recorded using BioSemi ActiveTwo System with a sampling rate of 2048Hz and stored in BioSemi Data Format (BDF) using ActiView BioSemi acquisition software. Each subject took approximately 20 minutes individually to complete an experiment.

The subjects were screened to select EEG data for data analysis and processing. The screening was based on SAM; subjects with low valence and arousal rating were rejected. The reason for screening was to select the most valuable data and remove the rest to get reliable results. The screening left 15 subjects out of 20. Screening was further applied to EEG data of 15 subjects to select the signal duration which fulfill the aimed emotion based on SAM. The idea behind this was to screen out and separate the data for each emotion. For example the signal for positive arousal were screened from the rest of the emotions and so on. EDF Browser[1] (a tool for reading and processing sensor data) was used to reduce the signals individually for the required duration. While reducing the signals, the first and last second had been eliminated from the total duration of five second stimulus presentations. This was in order to narrow down to exactly required data. The reason for this step was to focus on valuable data and filtering out the extra. Because when a picture is displayed, it takes some time for the brain to react to new stimuli and therefore the first second is usually noisy. Similarly, after looking at a picture stimulus for a while the brain goes into a relaxed state and does not react in the same activation as initially; therefore the last second was removed as well. This process was completed for pictures with positive, negative and neutral arousal as well as for positive, negative and neutral valence.

The screened data was preprocessed using EEGLAB Toolbox[2] for MATLAB. Epoch and Event info were extracted, the data was pruned and baseline removed. Finally, Independent Component Analysis (ICA) was performed on the data[32]. Preprocessing data with these various techniques helps to remove the artifacts such as eye blinking etc. This also make it easier to extract features from the signals.

[1] http://www.teuniz.net/edfbrowser
[2] http://sccn.ucsd.edu/eeglab

Feature selection is one of the key challenges in affective computing due to phenomena of person stereotype[13]. This is because different individuals express the same emotion with different characteristic response patterns for the same situations. Each subject involved in the experiment was having diverse physiological indices that showed high correlation with each affective state. The same finding has been observed by Chen and Hou[25] and is explained by Rani et al.[33]. From the obtained EEG data, it was observed that physiological features were highly correlated with the state of arousal among two subjects. According to Rani et al., a feature can be considered significant and selected as an input to a classifier if absolute correlation is greater for physiological features among subjects[33]. Based on these findings, it was observed that the accuracy improved for some techniques (i.e. KNN, BNT and ANN) when highly correlated features were used, while it degraded for the others (i.e. RT and SVM). Chen and Hou point out that selection of highly correlated features helps to exclude the less important features for affective state and hence improve the results[25].

The preprocessed data was further processed to get the real values for the signals using EEGLAB Toolbox for MATLAB. Based on findings by AlZoubi et al. the four features *minimum value*, *maximum value*, *mean value* and *standard deviation* were extracted from each signal in order to further process the data[31].

The raw EEG data is processed to extract the selected features. Different signal processing techniques are available for this purpose such as Fourier transform, wavelet transform, thresholding, and peak detection. The values obtained were formatted in Attribute-Relation File Format (ARFF), which is an acceptable file format for the data-mining tool WEKA[3]. The values obtained are used as instances in the ARFF file with a binary class value as negative and positive arousal/valence. Each feature value (min value, max value, mean value and standard deviation) for each electrode is a separate attribute in each instance in the ARFF file. Six electrodes were used making the total number of attributes 24 (plus the class value). A separate dataset was created for each subject, as well as a combined dataset with data from all subjects.

Each dataset were classified using machine learning techniques available in WEKA. During the classification, the classifier was trained to classify negative or positive arousal/ valence values as correctly classified whereas neutral values as incorrectly classified. The techniques used had all the default parameter values as implemented in WEKA. In all experiments 10 fold cross validation were used.

Figure 1 shows the complete process of capturing, processing and classifying the EEG data in the conducted experiments. The results from classifying the EEG data for all 15 subjects are presented in Table 1 and Figure 2. The highest accuracy was obtained with SVM (56.10%) closely followed by KNN, RT and BT (52.44%). The three latter all had the same accuracy indicating that they at least in this case discriminate the data in a similar way. The result are not very promising indicating that there can still be noise in the processed data, or that the selected features are not representative for all subjects which can be a problem as pointed out by Rani et al.[13]. As comparison a random guess would give an accuracy of 33% since three possible emotional states (positive or negative valence/arousal and neutral) are used.

[3] http://www.cs.waikato.ac.nz/ml/weka

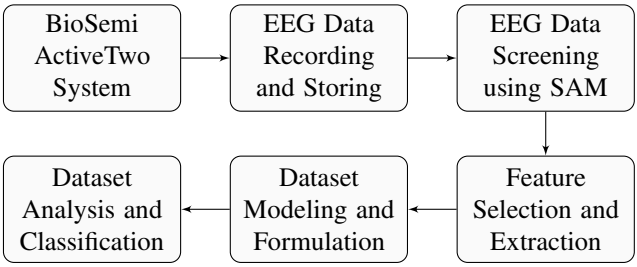


Fig. 1. The process for capturing, processing and classifying the EEG data

Table 1. Results from classifying EEG data for all subjects

Technique	Accuracy
K-Nearest Neighbor	52.44%
Regression Tree	52.44%
Bayesian Network	52.44%
Support Vector Machine	56.10%
Artificial Neural Networks	48.78%
Random guess	33.33%

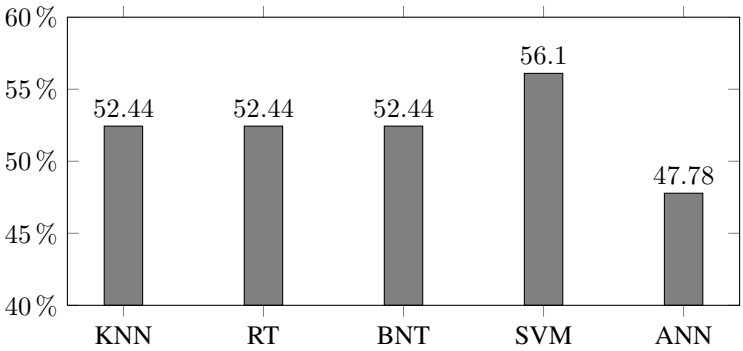


Fig. 2. Results from classifying EEG data for all subjects

To see if there could be problems with the generality of the selected features we divided the dataset into three subsets each with data from five subjects. The subsets were split in a semi-random fashion. The first five subjects was put in Dataset 1, the next five in Dataset 2 and the last five in Dataset 3. The results are shown in Table 2 and Figure 3. They show that all classifiers except RT had difficulties classifying Dataset 3. RT had problems classifying both Dataset 2 and 3. In this experiment SVM are still

the best classifier followed by KNN. It is interesting that KNN, RT and BN all had the same accuracy when classifying the full dataset, but in this case RT and BN are well behind KNN. The best results (over 70% accuracy topping at 77.78%) are in line with the accuracy of other related experiments (see for example [2]).

Table 2. Results from classifying datasets of five subjects each

Technique	Dataset 1	Dataset 2	Dataset 3	Average
K-Nearest Neighbor	70.37%	66.67%	51.35%	62.80%
Regression Tree	62.96%	44.44%	45.95%	51.12%
Bayesian Network	59.26%	55.44%	48.65%	54.45%
Support Vector Machine	77.78%	70.27%	51.35%	66.47%
Artificial Neural Networks	70.37%	61.11%	43.24%	58.24%
Random guess		33.33%		

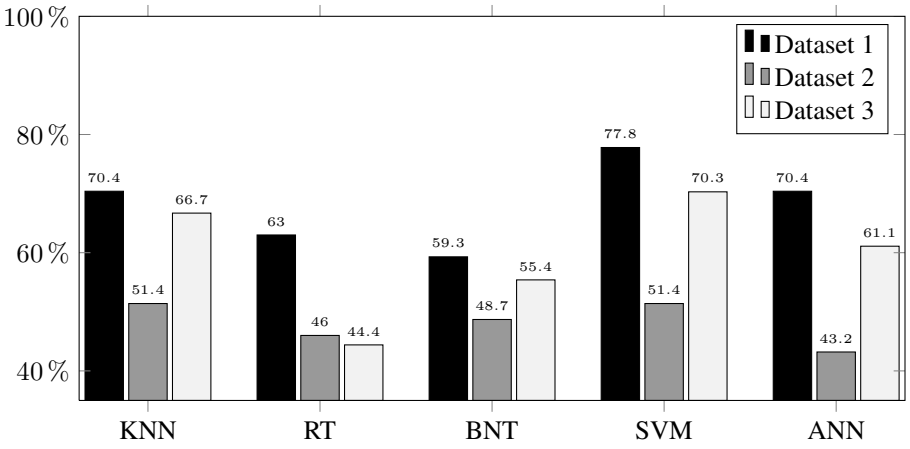


Fig. 3. Results from classifying datasets of five subjects each

In the last experiment we used datasets containing of only a single subject. This was done for the first three subjects. The results are shown in Table 3 and Figure 4. In this experiment KNN was the most accurate classifier with 83.33% accuracy for Subject 3. It is interesting to see that SVM was only able to get 50.00% accuracy on the same subject. BN showed very large differences with 72.72% accuracy for Subject 2 and only 36.36% for Subject 1.

Table 3. Results from classifying datasets of single subjects

Technique	Subject 1	Subject 2	Subject 3
K-Nearest Neighbor	54.54%	72.72%	83.33%
Regression Tree	36.36%	54.54%	50.00%
Bayesian Network	36.36%	72.72%	66.66%
Support Vector Machine	45.45%	45.45%	50.00%
Artificial Neural Networks	45.45%	45.45%	50.00%
Random guess		33.33%	

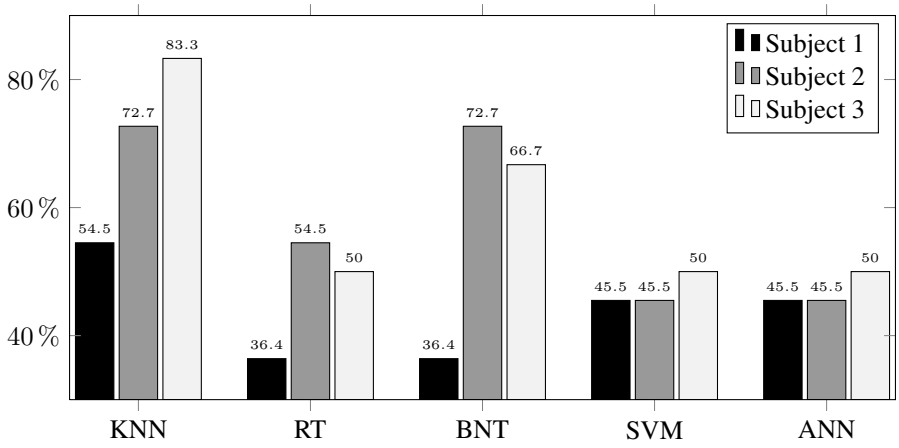


Fig. 4. Results from classifying datasets of single subjects

3 Discussion and Future Work

The main purpose of our experiments was to evaluate different machine learning techniques for classifying EEG data. From our results we can conclude that it is not trivial to process and classify data to be accurate over a large number of subjects. The results from all 15 participants was in the best case 56.10%. When dividing the subset into three parts with five subjects each the accuracy rose to 77.78%. In both cases SVM was the best classifier with KNN slightly behind. The results from classifying data from single subjects showed an accuracy of 83.33% for KNN. Interesting is that SVM only showed an accuracy of 50.00% on single subjects.

As Rani et al. discusses the feature selection is a key challenge in affective computing due to phenomena of person stereotype[13]. This is probably the reason why the accuracy in our experiments greatly increased on smaller datasets. It is difficult to find features that are generally working well over a large number of subjects. Another reason is that EEG data is noisy and diverse and is often very difficult to work with. There is also the possibility that the IAPS pictures did not induce strong enough emotions on some subjects making it difficult to classify some emotional states.

Based on the results we cannot say which classifier that generally is the best, but KNN and SVM seems to be good choices regardless of the size of the datasets.

In the future we would be interested in using more features and different combinations of them to see how it affects the accuracy over many subjects. It would also be interesting to observe if more subjects in the experiment would have any positive or negative impact on the results, as the amount of data for the classifier increases. In this experiments we used a binary class value for the classifiers (negative or positive valence/arousal) and an unknown as neutral valence/arousal. It could have impact on the results if we use three separate classes with neutral valence/arousal as its own class value instead.

References

1. Liu, Y., Sourina, O., Nguyen, M.K.: Real-Time EEG-Based Human Emotion. In: Proceedings of the 2010 International Conference on Recognition and Visualization (2010)
2. Petrantonakis, P.C., Hadjileontiadis, L.J.: Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis. IEEE Transactions on Affective Computing 1, 81–97 (2010)
3. Davidson, R.J., Ekman, P., Saron, C.D., Senulis, J.A., Friesen, W.V.: Withdrawal and cerebral asymmetry: Emotional expression and brain physiology. Journal of Personality and Social Psychology 58, 330–341 (1990)
4. Fisher, R.: The use of multiple measurements in taxonomic problems. Annals of Human Genetics 7, 179–188 (2008)
5. Efron, B.: Least angle regression. Ann. Statist. 32, 407–499 (1997)
6. Jaakkola, T., Jordan, M.: A variational approach to Bayesian logistic regression models and their extensions. In: Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics (2008)
7. Wilson, G.F., Russell, C.A., Monnin, J.W., Estepp, J.R., Christensen, J.C.: How Does Day-to-Day Variability in Psychophysiological Data Affect Classifier Accuracy? In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2010)
8. Millan, J.R., Renkens, F., Mourino, J., Gerstner, W.: Noninvasive brain-actuated control of a mobile robot by human EEG. IEEE Transactions on Biomedical Engineering 51, 1026–1033 (2004)
9. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1175–1191 (2001)
10. Nasoz, F., Alvarez, K., Lisetti, C.L., Finkelstein, N.: Emotion recognition from physiological signals for presence technologies. International Journal of Cognition, Technology, and Work 6 (2003)
11. Conati, C.: Probabilistic assessment of users emotions in educational games. Applied Artificial Intelligence 16, 555–575 (2002)
12. Wilson, G.F., Russell, C.A.: Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. Human Factors: The Journal of the Human Factors and Ergonomics Society 45, 635–644 (2003)
13. Rani, P., Liu, C., Sarkar, N., Vanman, E.: An empirical study of machine learning techniques for affect recognition in human robot interaction. Pattern Analysis and Applications 9, 58–69 (2006)

14. Lin, Y., Wang, C., Wu, T., Jeng, S., Chen, J.: EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2009)
15. Bos, D.: EEG-based Emotion Recognition. The Influence of Visual and Auditory Stimuli, pp. 1–17 (2006)
16. Horlings, R., Datcu, D., Rothkrantz, L.J.M.: Emotion recognition using brain activity. In: Proceedings of the 9th International Conference on Computer Systems and Technologies (2008)
17. Murugappan, M., Rizon, M., Nagarajan, R., Yaacob, S., Zunaidi, I., Hazry, D.: Lifting scheme for human emotion recognition using EEG. In: Proceedings of the International Symposium on Information Technology (ITSim) (2008)
18. Schaaff, K.: EEG-based Emotion Recognition. Diplomarbeit am Institut fur Algorithmen und Kognitive Systeme. Universitat Karlsruhe (2008)
19. Li, M., Chai, Q., Kaixiang, T., Wahab, A., Abut, H.: Eeg emotion recognition system. In: In-Vehicle Corpus and Signal Processing for Driver Behavior, pp. 125–135 (2009)
20. Yu, Y., Lai, P., Ko, L., Chuang, C., Kuo, B., Lin, C.: An EEG-based classification system of Passengers motion sickness level by using feature extraction/selection technologies. In: Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN) (2010)
21. Parvin, H., Alizadeh, H., Minaei-Bidgoli, B.: MKNN: Modified k-nearest neighbor. In: Proceedings of the World Congress on Engineering and Computer Science (WCECS) (2008)
22. Downey, S., Russell, M.J.: A Decision Tree Approach to Task Independent Speech Recognition. In: Proceedings of the Inst Acoustics Autumn Conf on Speech and Hearing (1992)
23. Brown, L.E., Tsamardinos, I., Aliferis, C.F.: A novel algorithm for scalable and accurate bayesian network learning. In: Proceedings of the 11th World Congress on Medical Informatics (MEDINFO) (1992)
24. Macas, M., Vavrecka, M., Gerla, V., Lhotska, L.: Classification of the emotional states based on the EEG signal processing. In: Proceedings of the 9th International Conference in Information Technology and Applications in Biomedicine (2009)
25. Chen, G., Hou, R.: A New Machine Double-Layer Learning Method and Its Application in non-Linear Time Series Forecasting. In: Proceedings of the 2010 International Conference on Mechatronics and Automation (ICMA) (2007)
26. Huang, W.Y., Shen, X.Q., Wu, Q.: Classify the number of EEG current sources using support vector machines. Machine Learning and Cybernetics (2002)
27. Tangermann, M., Winkler, I., Haufe, S., Blankertz, B.: Classification of artifactual ICA components. International Journal on Bioelectromagnetism 11, 110–114 (2009)
28. Ho, C.K., Sasaki, M.: EEG data classification with several mental tasks. In: Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics (2002)
29. Lang, P., Bradley, M., Cuthbert, B.: International affective picture system (iaps): Affective ratings of pictures and instruction manual. Technical report a-8.). University of Florida, Gainesville, FL, Tech. Rep.
30. Homan, R.W., Herman, J., Purdy, P.: Cerebral location of international 1020 system electrode placement. Electroencephalography and Clinical Neurophysiology 66, 376–382 (1987)
31. AlZoubi, O., Calvo, R.A., Stevens, R.H.: Classification of EEG for Affect Recognition: An Adaptive Approach. In: Nicholson, A., Li, X. (eds.) AI 2009. LNCS, vol. 5866, pp. 52–61. Springer, Heidelberg (2009)
32. Ungureanu, M., Bigan, C., Strungaru, R., Lazarescu, V.: Independent component analysis applied in biomedical signal processing. Measurement Science Review 4 (2004)
33. Rani, P., Sarkar, N., Smith, C.A., Kirby, L.D.: Anxiety detecting robotic system towards implicit human-robot collaboration. Robotica 22, 85–95 (2004)

From Explicit to Implicit Speech Recognition

Chad M. Spooner[1], Erik Viirre[2], and Bradley Chase[3]

[1] NorthWest Research Associates Monterey, CA
cmspooner@nwra.com
[2] Department of Neurosciences UCSD, San Diego, CA
eviirre@ucsd.edu
[3] Department of Industrial and Systems Engineering USD, San Diego, CA
bchase@sandiego.edu

Abstract. We consider the problem of determining the word or concept that a subject holds in their mind prior to the act of speech using only a scalp-recorded electroencephalogram (EEG). Such speech acts are called covert, silent, or implicit speech acts in the literature. We consider a binary-tree classifier that uses one of a number of candidate feature types, including temporal correlation coefficients, spectral correlation, and time-gated raw voltages. The particular features and binary-tree parameters are blindly determined using the local discriminant basis (LDB) technique. The experiments involve sequential presentation of words and numbers on a computer screen. The subject wears an EEG scalp cap and is instructed to first consider the stimulus, then speak it. Later, the subject is instructed to perform the same task without the actual utterance, resulting in implicit speech. We present performance results for the various obtained classifiers, which show that the approach has significant merit.

1 Introduction

Speech recognition is a critical element of human-to-machine interfaces for an increasing number of applications. It is used for a variety of purposes from command and control to transcription. Most speech recognition applications begin with a general database of models, build extensive libraries of speaker-specific templates, and use Bayesian networks or other statistical means to apply word and grammar logic for more accurate interpretation. However, these applications remain challenged by inter-speaker variances, generalization to populations of speakers, noisy environments, and the ambiguities created by homophones and confusables. In addition, overt speech recognition applications are ineffective for many aphasias such as speakers who are impaired by stroke or traumatic brain injury, and locked-in subjects—individuals who have lost the ability to generate overt speech while retaining most or all other cognitive functions. These injured and locked-in subjects have few options for human-to-machine interfaces and often suffer with imperfect or non-existent human-to-human interaction. In response, researchers are studying the use of electromyograms (EMG) for subvocal

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 502–511, 2013.

speech recognition (cf. [1]). Subvocal speech is silent speech or what we call *implicit speech*. Subjects either mouth the words, or merely think about the words, but they do not vocalize the speech. For individuals in noisy environments or ALS patients who often can mouth words but cannot expel sufficient breath to vocalize, this technique holds promise.

However, EMG-based subvocal speech recognition has two major challenges. First, movement in the system severely disrupts attempts to match EMG patterns to intended speech. Second, the patterns thus far cannot be generalized—they are individual specific. In parallel, beginning in 1975, researchers began testing the feasibility of engineering two-way human-to-machine interfaces driven by the associated electrical activity of the brain. In [6], the authors concluded that both overt and implicit speech (i.e., thinking words without the corresponding overt utterance) recognition from EEG was possible if noise and interference sources could be minimized. Suppes and his collaborators demonstrated the ability to recognize one of seven words by first creating templates of averaged, simultaneously recorded EEG and MEG signals [14]. More recently, Viirre and Jung in [15] proved the ability to distinguish between spoken homophones by analyzing specific EEG components produced by an independent component analysis approach. In all cases, the authors illustrated that EEG activity during explicit speech was distinct and word-specific for a finite set of words and small numbers of subjects. Pinneo and Hall [6] specifically investigated the EEG of implicit speech and concluded that it was also distinct and word-specific.

Neural Basis for Speech Production

Evidence suggests that specific brain processes are invoked in speech production and those processes provide related detectable and exploitable signals. The current theory of speech production is the byproduct of modern language perception and production theory. Driven primarily by study of aphasias, the theory provides a modern framework consisting of three large interacting systems: the language implementation system, the mediation system, and the conceptual system. These systems encompass large portions of the left hemisphere in 96% of the population.

From [5], fMRI results from a subject instructed to utter a reveal successive activation in the cerebellum, basal ganglia, thalamus, cingulated motor area, primary motor cortex, and the supplementary motor area. The activation, deactivation, and interaction of these widely varying regions during lexical selection (cf. [2]), and syntactics and semantics (cf. [3]) processing are therefore potential sources of signals for implicit speech recognition.

2 Technical Approach: Advanced Signal Processing Techniques Applied to EEG

The distributed nature of neural activity naturally leads to an assumption that the spatial relationships of that activity contain exploitable information. Past work performed during the Advanced Signal Processing for Neuroscience project

(ASPN) [13] supports this supposition. We investigate the utility of several signal processing tools to the problems of characterizing and exploiting temporal and spatial EEG signal variations during implicit speech production.

Correlation and Temporal Coherence.
The spatial component of the EEG signal suggests that the cross correlation function and the normalized cross-correlation coefficient could be excellent classification measures. When combined with high spatial-resolution EEG recordings, both measures exploit the full breadth of space and time information. Consider two discrete-time signals $x(t)$ and $y(t)$. The cross correlation is defined by

$$R_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} x(t)y^*(t-\tau),$$

The correlation coefficient (temporal coherence) is the normalized cross correlation. The normalizing coefficient is the geometric mean of the mean-square of the two involved signals,

$$C_{xy}(\tau) = \frac{R_{xy}(\tau)}{[R_x(0)R_y(0)]^{1/2}},$$

where $R_x(\tau)$ is the autocorrelation function for $x(t)$. Features resulting from the examination of the correlation and coherence of EEG data across space and time provide task-specific indications of time-phased neural activity associated with specific experiment tasks. By exploiting these features it is possible to distinguish subtle differences in tasks that have been indistinguishable using other more traditional methods.

Cyclostationary Signal Processing.
Another statistical signal processing approach shown to reveal new information in scalp-recorded EEG signals arises from modeling EEG as cyclostationary signals [7,8]. Cyclostationary signals are the result of non-linear mixing between a stationary signal component and periodic signal components. The resulting signal has statistics that vary periodically with time, in many cases with multiple fundamental periods (polycyclostationary). In communications, the mixing of a sinusoidal carrier wave with multiple nonstationary components such as the information bit waveform or framing, results in a polycyclostationary signal. The same general mixing processes occur within the complex environment of neural activity and can manifest in exploitable cyclostationarity in EEG recordings. Two key functionals that are indicative of the nature and degree of cyclostationarity are the spectral correlation function and the normalized spectral coherence. The temporal correlation between two narrowband frequency components of a signal can be nonzero for nonstationary signals (i.e., cyclostationary signals). This is called spectral correlation and is defined in terms of a generalization of the power spectrum,

$$S^{\alpha}_{x_{T,\Delta f}}(f) = g_{\Delta f}(f) \otimes \left[\frac{1}{T} X_T(f+\alpha/2) X^*_T(f-\alpha/2)\right],$$

where $g_{\Delta f}(f)$ is a pulse-like smoothing function with width Δf , $\otimes$ denotes convolution, and $X_T(f)$ is the finite-time Fourier transform for $x(t)$. The ideal spectral correlation function is obtained by the following double limit

$$S_x^\alpha(f) = \lim_{\Delta f \to 0} \lim_{T \to \infty} S_{x_{T,\Delta f}}(f)$$

The spectral correlation function can be converted to a correlation coefficient by normalizing by the geometric mean of the variances of the two quantities involved in the correlation operation. This leads to the spectral coherence function,

$$C_x^\alpha(f) = \frac{S_x^\alpha(f)}{[S_x^0(f+\alpha/2)S_x^0(f-\alpha/2)]},$$

where $S_x^0(f)$ is the power spectrum for $x(t)$. Features derived from the spectral correlation and spectral coherence functions of scalp-recorded EEG signals could reveal rich and potentially exploitable structure.

Wavelets and the Local Discriminant Basis.

The local discriminant basis (LDB) [9,10] is another complex modern signal processing tool aimed at automatic identification of powerful wavelet-based discrimination functions for use in arbitrary M-class signal- or image-classification problems. The explanation below is adapted from [11] closely except where noted.

The wavelet transform of an $N \times N$ image is defined by a pair of quadrature mirror filters, h and g and a maximum decomposition depth D [4]. The filters h and g are low- and highpass filters, respectively. The transform applies the filters to the rows and columns of the image interactively, subsamples the results, and begins the process anew with the subsampled data. The filters are applied in all four of their row-column combinations: low-low (LL), low-high (LH), high-low (HL), and high-high (HH), resulting in L_x, V_x , H_x, and D_x residuals. At each iteration, the convolution-sampling operation is applied to the L_x data only, while the other three data sets are retained as is. At the final stage (stage D) the L_x coefficients are also retained.

Now, suppose the context of the wavelet decomposition involves a classification problem with C classes, and we have N_c training images for classes $c = 1, \ldots, C$, provided in sets X_c. The LDB finds a basis such that there are a few basis vectors whose coefficients vary widely among the classes, while varying little between members of a class. Thus, the basis provides an excellent tool for classification. Before selecting a vector for good discrimination between classes, a measure of the average strength of the vector coefficients is required. Then by establishing a measure of the distance or difference between the basis coefficients for two or more classes, this distance indicates the discrimination power of the corresponding basis vector. By using an algorithm to optimize the selection of the vectors, a final set of basis vectors results. The power of wavelet-based LDB lies in its ability to calculate the optimized basis vectors and classify data blindly. As different transforms are developed for EEG processing, the LDB tool calculates the basis vectors and performs classification without modification. This enables rapid evaluation of new transforms, features, and data sets.

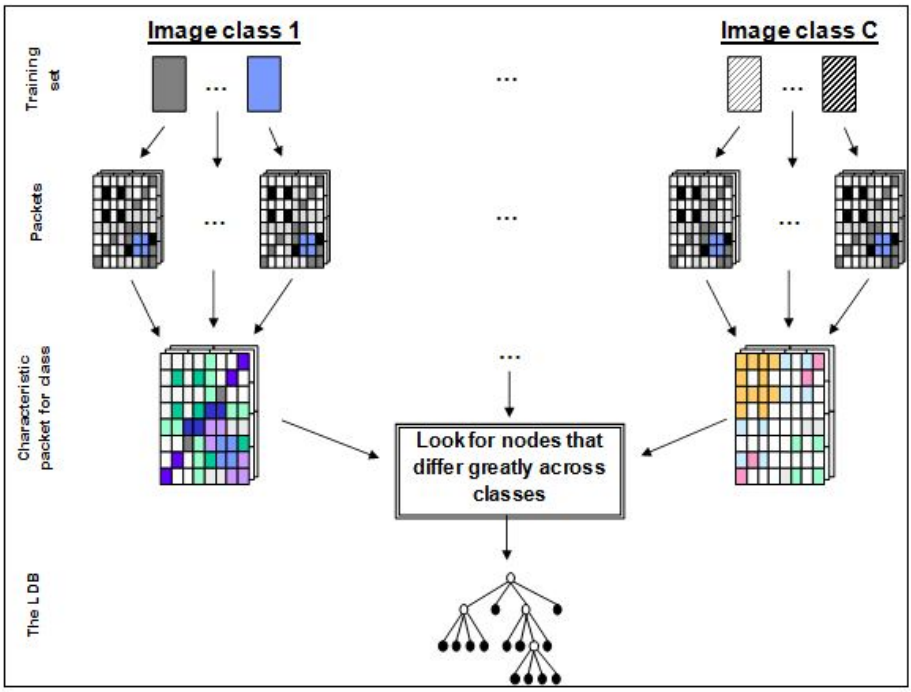


Fig. 1. The concept behind the local discriminant basis (LDB). The LDB is similar to the concept of *best basis* used in wavelet-based signal and image compression. Instead of finding the most parsimonious basis to represent an input, the LDB finds the basis that best discriminates between the classes involved in the classification problem.

3 Experimental Methods

The EEG procedure used was a standard clinical EEG collection that involves placing a stretchable cap containing 128 small electrodes on the subject's head and then filling each electrode with a water-soluble gel. In addition, electrodes were taped above and below one eye, as well as on the right and left to monitor eye movements. After the EEG systems were in place a baseline reading was obtained for 5 minutes. During baseline the subjects sat in a comfortable chair in front of a blank computer monitor. After the baseline was obtained the computer word task started. The task lasted approximately 30 minutes. Subjects were shown words individually and asked to speak clearly after the word appeared. Words appeared in random order in intervals of 5–7 seconds. Subjects were given words from sets such as *too* vs. *two* and *four* vs. *for*. Over a 30-minute session, each individual word was spoken 100 times. Subjects provided data for four separate experiments. The experiments were designed to progress from explicit speech recognition to implicit speech recognition with the final experiment designed to approximate real-world conditions. The experiments were:

1 **Explicit Speech Output.** Explicit speech utterances augmented by voice recording of the utterance, and electromyography (EMG) channels.
2 **Explicit Speech Expanded Set.** Explicit speech utterances for an expanded set of words and phrases augmented by voice recording of the utterance, and electromyography (EMG) channels.
3 **Implicit Speech.** Implicit speech "utterances" augmented by voice recording of the utterance, and electromyography (EMG) channels. The full set of words and numbers presented as stimuli for this experiment is: {1, 2, 4, 8, Ate, Eight, For, Left, One, Right, Two, Won}.

4 Results

Data was collected for all three experiments from 12 subjects. Data from one subject was corrupted and not processed. Data from the remaining 11 subjects was processed first using algorithms developed under the ASPN program. Subsequently, we focused on Experiment 3, the implicit speech task, processing that data with both ASPN algorithms, and new techniques developed specifically for the implicit speech data.

Because it most closely approximates the performance goals of an operational system and simulates a real-world environment, Experiment 3 produced the most critical data. Thus, the results from processing Experiment 3 data are highlighted below. Figure 2 shows the first of three views of the processing results. For each two-class problem derived from the twelve presented stimuli, (i.e., is the subject implicitly speaking Stimulus 1 or Stimulus 2?) we have plotted the maximum probability of correct classification (PCC) in the bar graph. Each of the stimuli are represented on the x-axis and the maximum two-class PCC is plotted on the y-axis. There are four colored bars for each stimulus. They represent the four feature sets we used to classify the trials. The blue bar is the performance using the temporal coherence ("Second-Order Statistics"), the green bar shows performance using cyclostationary statistics ("Cycle Frequencies"), the pink bar is performance using features derived directly from the filtered EEG data in the full two seconds for each trial. Finally, the black bar shows the performance when using the filtered EEG data features but restricting the time interval to between 200 and 800 milliseconds post stimulus. The solid red line shows a PCC of 50% which is the result for random guessing. In nearly every case, the time-restricted, filtered EEG data provided the maximum performance. All four feature sets show maximum performance superior to random guessing.

The second view of our implicit-speech experimental results is shown in Figure 3. Here, the four colored lines represent performance with the corresponding feature sets. In this plot, the relative frequency of occurrence for a given PCC is shown. The red dashed line is again the PCC for random guessing. When more area under a curve can be found to the right of the red dashed line than to the left, then our classifier performed better than random guessing for that particular two-class problem. Note again that the black line representing the performance of the restricted time interval features has the vast majority of area to the right of the random-guess line.

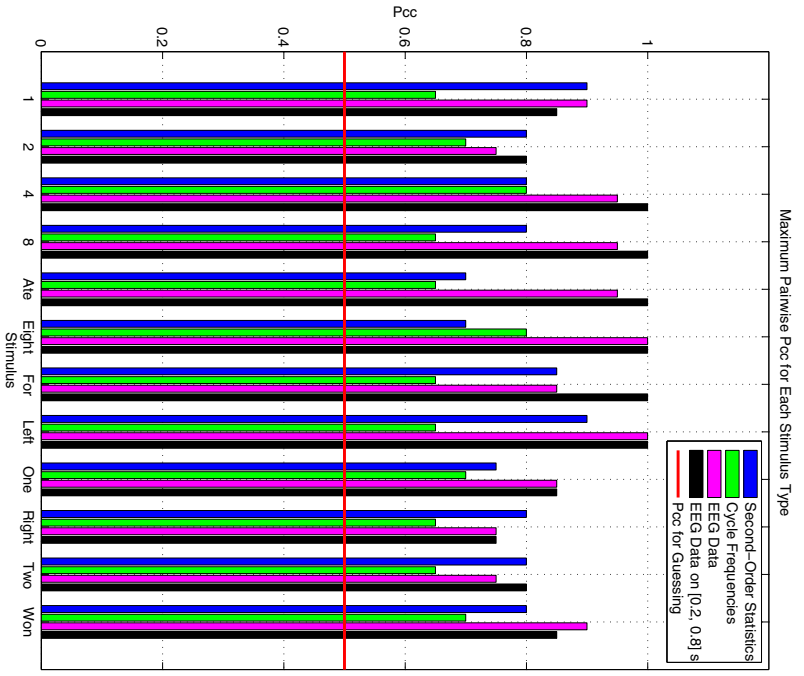


Fig. 2. Maximum PCCs for Subject 3 implicit-speech experiment

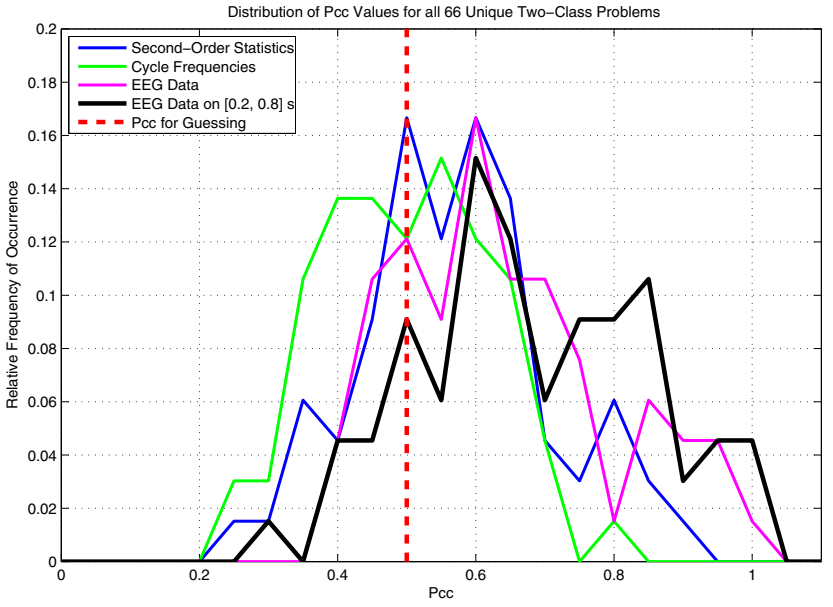
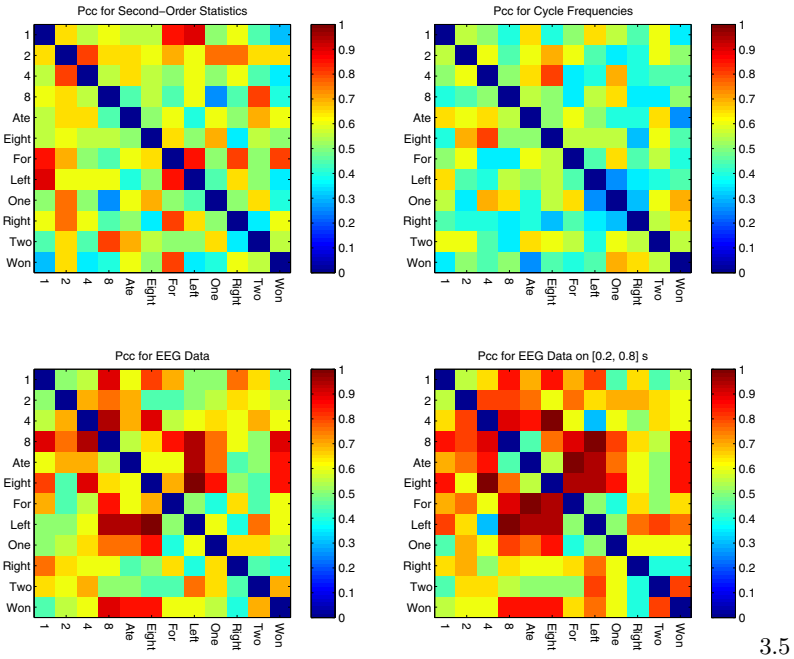


Fig. 3. PCC frequency-of-occurrence for Subject 3 implicit-speech experiment



3.5

Fig. 4. PCCs for Subject 3 implicit-speech experiment

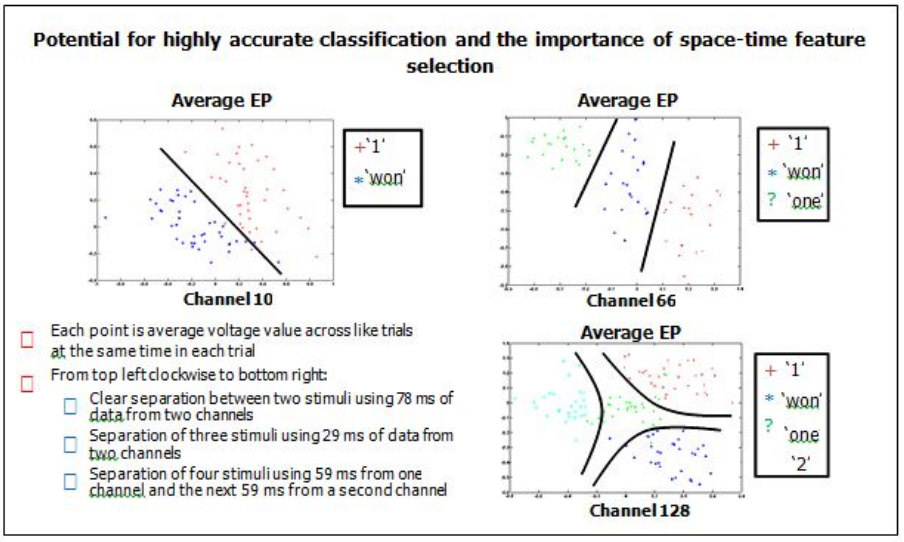


Fig. 5. Hand-culled voltages taken from various EEG channels for an explicit-speech experiment. Note the high degree of separability by stimulus type. This type of separability is consistent with our Subject-3 blind feature identification and classification results.

The third representation of our results is shown in Figure 4, in which the PCCs for all two-class problems and all classification-feature types are displayed as matrices. From this figure, it is clear that using the time-gated raw EEG voltages provides the best overall performance for Subject 3. However, the different feature types provide different performance levels for the various two-class problems, suggesting that we might obtain substantially better performance by properly combining the features in a new classifier.

Detailed EEG Voltage Analysis

In addition to the blind feature-identification and processing using the local discriminant basis, we also extracted voltages from various channels for several stimuli in an attempt to gain a qualitative understanding of the potential for feature separation by stimulus type. An example is shown in Figure 5.

5 Discussion and Conclusions

The processing results show a clear proof of concept for decoding implicit speech using only scalp-recorded EEG. The results were obtained by processing raw EEG signal data to approximate real-world processing. There were no steps taken to eliminate artifacts from eye blinks, head movement, or any other noise introduced into the signals. In addition, the analysis presented is independent of the signal environment. There were no brain models (signal/channel models) employed to develop the algorithms.

The real-world signal environment was approximated further by avoiding calibration or subject training routines and by specifically isolating the signals to EEG only and limiting the EEG to non-motor related EEG. No pre-motor or motor program content was used. Finally, no prior knowledge of the signals was assumed so the algorithms were specifically designed to operate blindly, exploiting unknown signals.

A significant issue that was uncovered and addressed involved the challenging real-world timing of the features used to decode the implicit utterances. The EEG data contains features for implicit utterances with loosely constrained timing meaning the utterance onset varies randomly over the trials as it would in an operational environment. The algorithms still identify signal patterns related to implied utterances independent of this random timing making these techniques tolerant to unknown delays.

The algorithms extracted more information from the frequency domain (exploited wider bandwidth and higher frequencies), from the spatial domain (analyzed all viable channels), and from the time domain (derived features from several hundred milliseconds) than any previous work in speech recognition. By using every degree of freedom available and narrowing the feature set to a small number of key features, the algorithms achieved results extendable to an operational system. Finally, the signal processing results correlate with neurolinguistics. Analysis of timing and channel locations are consistent with neurological functions associated with the production of an utterance. Because of this,

evidence suggests roughly ten (or less) specific EEG channels will be required for the ultimate applications of decoded implicit speech.

References

1. Chan, A., Englehart, K., Hudgins, B., Lovely, D.: Hidden Markov model classification of myolectric signals in speech. IEEE Engineering in Medicine and Biology Magazine 21, 143–146 (2002)
2. Gazzaniga, M. (ed.): The Cognitive Neurosciences. MIT Press, Cambridge (1995)
3. Kutas, M., Hillyard, S.A.: Event-related brain potentials to semantically inappropriate and surprisingly large words. Biol. Psychol. 11(2), 99–116 (1980)
4. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press, San Diego (1998)
5. Soros, P., et al.: Clustered functional MRI of overt speech production. Neuroimage 32(1), 376–387 (2006)
6. Pinneo, L.R., Hall, D.J.: Feasibility Study for Design of a Biocybernetic Communication System. Final Technical Report: ARPA Contract No, Stanford Research Institute, Menlo Park, CA (1975)
7. Gardner, W.A.: Exploitation of Spectral Redundancy in Cyclostationary Signals. IEEE Signal Processing Magazine, 14–36 (1991)
8. Spooner, C.M., Gardner, W.A.: The Cumulant Theory of Cyclostationary Time-Series, Part I: Foundation and Part II: Development and Applications. IEEE Trans. Sig. Proc. 42, 3387–3429 (1994)
9. Saito, N.: Local Feature Extraction and its Applications using a Library of Bases. Ph.D. Dissertation, Yale University (1994)
10. Saito, N., Coifman, R.R.: Improved Local Discriminant Bases Using Empirical Probability Density Estimation. In: Proceedings of the Joint Statistical Meeting, Chicago (1996)
11. Spooner, C.M.: Applications of Local Discriminant Bases to HRR-Based ATR. In: Proceedings of the Thirty-Fifth Annual Asilomar Conference on Signals, November 4-7 (2001)
12. Streight, D.A.: Application of higher-order wavelet decomposition statistics to blind steganalysis of still images, SDSI Technical Report (2005)
13. Streight, D.A., Spooner, C.M.: Final Report: Advanced Signal Processing for Neuroscience. DARPA/US Army Contract #NBCHC060079 (2007)
14. Suppes, P., Lu, Z., Han, B.: Brain wave recognition of words. Proc. Natl. Acad. Sci. 94, 14965–14969 (1997)
15. Viirre, E., Jung, T.-P.: Augmented Higher Cognition: Enhancing Speech Recognition Through Neural Activity Measures. In: Proceedings of the 11th Human Computer Interaction Conference, vol. 11-Foundations of Augmented Cognition (2005)

Cognitive-Affective Interactions in Strategic Decision Making

Yanlong Sun and Hongbin Wang

School of Biomedical Informatics, University of Texas Health Science Center at Houston, USA
{Yanlong.Sun,Hongbin.Wang}@uth.tmc.edu

Abstract. While making a decision to maximize the expected utility is among the prime examples of human intelligence, the ultimatum game showcases a social dilemma where people sacrifice their economic self-interest in the presence of negative emotions. In the present study, we explore human cognitive-affective interactions in strategic thinking from an integrated neurocomputational perspective. We manipulated participants' emotions by inducing incidental affective states in the ultimatum game. We found that participants' rejection rates of unfair offers were significantly lower in positive valence emotions ("happy" and "calm") than in negative valence emotions ("sad" and "anxious"). In addition, the reduction of rejection rates appeared to be independent of the arousal level (high arousal in "happy" and "anxious" versus low arousal in "calm" and "sad"). Our results suggested that positive valence emotions, by broadening people's evaluations of decision perspectives and alleviating the perception of unfairness, may help people regain focus on their economic self-interest.

Keywords: Decision making; social dilemma; ultimatum game; affective induction; fairness preference; valence; arousal.

1 Introduction

Normative theories of judgment and decision-making in economics typically assume people to be rational and self-regarding [e.g., 1]. However, it has been documented that in the context of social interactions, people do not always act to maximize their self-interest according to the utility functions. One prominent example is the ultimatum game, a relatively recent showcase of human "irrationality" in decision making [2]. In a simple form of the game, two players decide how to divide a $10 award. One player (the proposer) makes an offer and the other player (the responder) decides whether to accept the offer. If the responder accepts the offer, the award is split as proposed. If the responder rejects the offer, both players get nothing. Suppose that the proposer may make any offer from $0 through $10, presumably a "rational" (i.e., utility maximizing) responder should accept any non-zero offer, even if the offer is "unfair" (e.g., less than $5), since the alternative is getting nothing. The dominant empirical finding, however, is that the responder often rejects an offer less than 30% of the sum, a clear deviation from the prediction of normative theories [for a review, see 3].

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 512–520, 2013.

A straightforward explanation for the rejection behavior in the ultimatum game is that the players' decisions depend on not only their own payoffs but also their perception of fairness, and the rejections of unfair offers reflect people's preference of fairness-seeking [4-7]. Knoch and colleagues [8-10] suggest that self-interest and fairness preferences operate via different systems: self-interest is the more evolutionarily primitive desire but can be suppressed by the fairness preferences in order to enforce social norms. Instead of the fairness perception, other studies emphasize the role of emotions in the ultimatum game. The wounded pride/spite model [11] posits that responders perceive small offers as unfair, and therefore react with anger and spiteful rejections [also see 12]. Pillutla and Murnighan [13] find that rejections were most frequent when responders could evaluate the fairness of the offers and suggest that anger was a better explanation of the rejections than the perception of unfairness. Mikula, Scherer, and Athenstaedt [14] show that injustice elicits anger, disgust, sadness, and other negative emotions. Functional neuro-imaging studies have revealed that unfair offers induce activations in brain regions that are associated with disgust [15, 16].

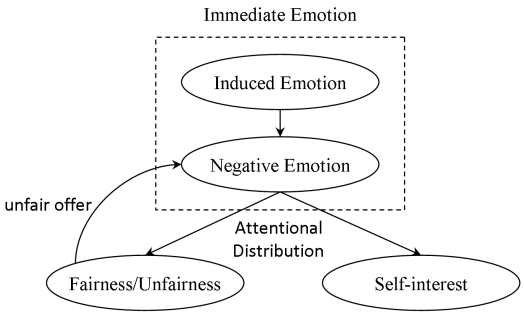


Fig. 1. An attentional distribution network in which perceptions of fairness and self-interest are modulated by the immediate emotion

From an integrated neurocomputational perspective, Wang, Coble, and Bello [17] propose that unfair offers in the ultimatum game lead to cognitive-affective interactions, in which the utility calculation in the posterior cortices is modulated by the affective states represented in the lower-level sub-cortical areas. This account points to a biologically realistic model in which emotional and cognitive processes are integrated into one attentional distribution network (Fig. 1). Specifically, we conjecture that in resolving the conflict between self-interest and fairness, emotions can have a causal effect on decision-making due to their roles in guiding attentional resources. When the players (respondents) consider an offer, they distribute their attentional resources between two preferences, self-interest and fairness, and such attentional distribution is regulated by the players' immediate emotion. When an offer is perceived as unfair, a conflict would arise because seeking for fairness now means rejecting the offer thus hurting self-interest. Because an unfair offer can elicit negative emotions [e.g., 11] and negative emotions tend to narrow the scope of people's

attention [e.g., 18], the players would be entrapped in the loop of focusing too much attention on the fairness preference, and consequently, ignore the aspect of self-interest.

In the present study, we examine whether emotion actually regulates the attentional distribution by inducing a range of incidental affective states (e.g., positive vs. negative valence, high vs. low arousal) that are independent of the fairness of offers. In the two experiments reported here, we manipulated the participants' immediate emotions with classic music clips [19, 20] as the main affect stimulus, enhanced by life event recall [19]. Our main focus was to compare participants' rejection rates in two sets of opposite emotional states: "happy" (positive valence and high arousal) versus "sad" (negative valence and low arousal) in Experiment 1, "calm" (positive valence and low arousal) versus "anxious" (negative valence and high arousal) in Experiment 2. Based on the documented functionalities of positive and negative emotions [19], we predict that compared with negative valence emotions ("sad" or "anxious"), positive valence emotions ("happy" or "calm") would make participants less distracted by the aspect of fairness thus focus more attention on self-interest, and consequently, lead to fewer rejections of unfair offers. Moreover, in dissociating valence and arousal, we speculate that the perception of fairness or unfairness would be more strongly associated with the valence than the arousal dimension of emotions.

2 Experiment 1

2.1 Participants

Seventeen participants (9 females and 8 males) participated in the experiment as responders, all of whom were graduate students or postdoctoral fellows in the Texas Medical Center (the mean age was 33.8 years with a standard deviation of 8.79 years).

2.2 Stimuli for Affective Inductions

Ten classical music clips from 18th, 19th, and 20th century Western composers were selected, five for each of the "happy" and "sad" affective inductions. These clips have been empirically validated to induce the corresponding affective states [20]. To enhance the inductions, we instructed the participants to silently recall into details of a happy or sad life event while listening to the music. Life-event recall, combined with music, has been used to successfully induce affect [19]. The standard ultimatum game involves only gains. Wang et al. [17] add a loss framing, in which the proposer and the responder split a cost of $10, and rejecting a proposal means both players having to each pay $10. It is possible that people's immediate emotion could interact with the perception of gain or loss. For example, one might feel "happier" considering a potential gain than considering a potential loss. For this reason, we adopted this two-frame game in the current experiment.

2.3 Design and Procedures

We used a 2 (affect conditions: "happy" and "sad") x 2 (framings: gain and loss) x 11 (offer amounts: $0, $1, $2…, $10) within-subject design. For each participant, the trials were grouped into 4 blocks (2 affect conditions x 2 framing conditions), and the orders of framing and affective conditions were counter-balanced between subjects. Within each block, each level of offer amount was repeated 3 times, resulting in 11 x 3 = 33 trials, and the order of trials was randomly shuffled.

The experiment was programmed in E-Prime and conducted on a PC with a 20 inch LCD monitor. After giving informed consent, participants were given instructions and practices of the game. They were told that they would play against individual anonymous proposers from a large online network, a new proposer for each game. At the beginning of each block, participants were instructed to develop a particular mood by listening to the music clips through the headsets for 5 minutes, followed by silently recalling in detail mood-appropriate events from their past. They were then instructed to rate their mood on a 9x9 grid by selecting a square that best exemplified their current mood in terms of valence (from "extremely sad" on the left to "extremely happy" on the right) and arousal (from "extremely low energy" at the bottom to "extremely high energy" at the top).

At the beginning of each game trial, participants were first prompted with a screen stating "New round! Connecting to a new partner …" for 2 seconds. This was to emphasize that each trial was a one-shot game with a different proposer such that the factor of reputation should not play a role here. In other words, rejecting the offer in the current trial would not serve as the means of punishing a unfair proposer in the previous trial. Then, depending on the framing condition, either "You get" or "You lose" was displayed for 1 second, which was followed by the amount of offer. Participants made a response by clicking either one of the mouse buttons to accept (left button) or reject (right button) the offer.

2.4 Experiment 1 Result

All 17 participants' data were included in data analyses. To examine whether affective inductions were effective, we first checked participants' self-reported ratings on emotional valence and arousal. Both ratings corresponded well to the intended affective states (see Table 1 and Figure 2). Compared with the "sad" condition, the "happy" condition resulted in higher ratings on both valence (mean difference = 4.03, paired $t(16) = 8.63$, $p < .01$) and arousal (mean difference = 2.53, paired $t(16) = 6.01$, $p < .01$).

On participants' rejection rates, we first examined the effects of affective conditions ("happy" vs. "sad") and framing domains (gain vs. loss) by repeated-measure ANOVA. Overall, affective conditions had a significant effect. Combining the corresponding columns in Table 2 (Experiment 1), it reveals that the overall rejection rate in the "happy" condition (22.9%) was significantly lower than in the "sad" condition (32.0%), with a mean difference of 9.1% ($F(1,16) \approx 7.03$, $p < .05$).

Table 1. Mean ratings on valence and arousal under each affective condition. Both ratings are scored in the range of [–4, 4] with 0 being neutral. Standard errors (over 17 participants in Experiment 1 and 12 participants in Experiment 2) are listed in parentheses. Column "Difference" is the absolute mean difference of valence or arousal ratings between "Sad" and "Happy" (Experiment 1), or, "Anxious" and "Calm" (Experiment 2), respectively. **: paired t-test, $p < .01$; *: $p < .05$.

Exepriment 1

	Happy	Sad	Difference
Valence rating	2.29 (0.27)	– 1.74 (0.35)	4.03 **
Arousal rating	1.82 (0.31)	– 0.71 (0.38)	2.53 **

Experiment 2

	Calm	Anxious	Difference
Valence rating	1.75 (0.26)	– 0.67 (0.61)	2.42 **
Arousal rating	– 0.58 (0.49)	0.88 (0.42)	– 1.46 *

Table 2. Mean rejection rates in percentage under each affect and framing conditions. Standard errors (over 17 participants in Experiment 1 and 12 participants in Experiment 2) are listed in parentheses. The bottom row lists the difference in rejection rates between "happy" and "sad" (Experiment 1), and, between "calm" and "anxious" (Experiment 2). In each framing condition, offers are split into sub-columns depending on whether they are less or greater than \$5: offers less than \$5 in the gain domain and greater than \$5 in the loss domain are considered "unfair" (in bold fonts).

Experiment 1

	Gain		Loss	
	< \$5 (unfair)	> \$5	< \$5	**> \$5 (unfair)**
Happy	**51.0 (7.2)**	3.9 (1.9)	0 (0)	**45.1 (7.1)**
Sad	**67.5 (7.8)**	4.3 (2.9)	2.4 (1.6)	**62.7 (8.8)**
Diff.	**16.5**	0.4	2.4	**17.6**

Experiment 2

	Gain		Loss	
	< \$5 (unfair)	> \$5	< \$5	**> \$5 (unfair)**
Happy	**52.8 (9.6)**	0.6 (0.6)	5.6 (4.4)	**42.2 (10.2)**
Sad	**60.6 (7.9)**	2.2 (1.7)	13.3 (6.9)	**53.3 (9.7)**
Diff.	**7.8**	1.6	7.7	**11.1**

The effect of framing and its interaction with the affective conditions were not significant ($F(1,16) \approx 1.90$, $p \approx .19$; $F(1,16) \approx 0.07$, $p \approx .80$, respectively). In addition, Table 2 (Experiment 1) shows that between the gain and loss domains, the rejections rates were almost symmetrically distributed across affective conditions. Since we

were particularly interested in whether the induced emotional states would alter participants' perception of fairness or unfairness, we separately examined two situations in which an offer was either "unfair" (less than $5 in the gain domain and greater than $5 in the loss domain) or "more-than-fair" (greater than $5 in the gain domain and less than $5 in the loss domain). The effect of affective conditions was statistically significant for "unfair" offers ($F(1, 16) \approx 10.30$, $p < .01$) but was not statistically significant for "more-than-fair" offers ($F(1, 16) \approx 0.36$, $p \approx .56$). The last row in Table 1 (Experiment 1) shows that the difference in the rejection rates between two affective conditions was always in the same direction across all columns (lower rejection rates in "happy" than in "sad"), but the magnitude was the greatest for unfair offers in both framing domains.

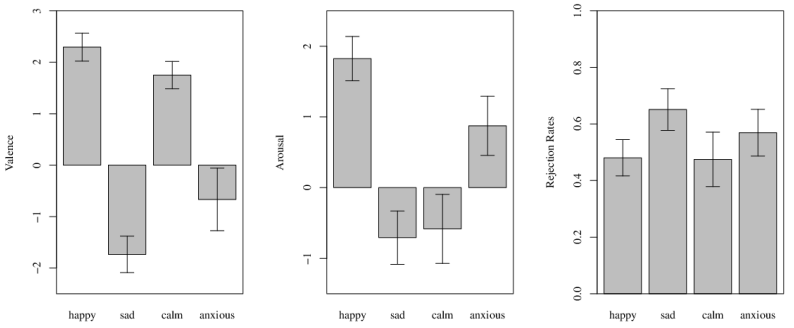


Fig. 2. Self-reported ratings on affective valence and arousal and rejection rates for unfair offers (gain and loss combined) in comparing "happy vs. sad" (Experiment 1) and "calm vs. anxious" (Experiment 2). Error bars represent one standard error above and one standard error below the mean.

3 Experiment 2

To dissociate the two dimensions of emotion, valence and arousal, we conducted the second experiment in which we compared two different affective states, "calm" (positive valence and low arousal) and "anxious" (negative valence and high arousal).

3.1 Participants and Procedure

Twelve participants (6 females and 6 males) who were not included in Experiment 1 participated in Experiment 2, all of whom were graduate students or postdoctoral fellows in the Texas Medical Center (the mean age was 35.5 years with a standard deviation of 7.44 years). Experiment 2 followed the same design and procedure as Experiment, except that we used classical music clips from [19], combined with life-event recall, to induce "calm" and "anxious" affective states.

3.2 Experiment 2 Result

All 12 participants' data were included in data analyses. Again, participants' self-reported ratings on emotional valence and arousal corresponded well to the intended affective states (see Table 1 and Figure 2). Compared with the "anxious" condition, the "calm" condition resulted in higher ratings on valence (mean difference = 2.42, paired $t(11) = 3.87$, $p < .01$) but lower ratings on arousal (mean difference = – 1.46, paired $t(11) = – 2.59$, $p < .05$). Compared with Experiment 1, the differences on both valence and arousal ratings between the two target emotional states were in smaller magnitudes. Nevertheless, in terms of dissociating valence and arousal, we have obtained an obvious contrast: Figure 2 shows that in contrast to the "happy-sad" comparison, the "calm-anxious" comparison was in the same direction on valence ratings, but in the opposite direction on arousal ratings.

Comparing two experiments, despite the reversed contrast on arousal ratings, rejection rates were similar between "happy" and "calm", and between "sad" and "anxious", respectively (both between-subjects comparisons were not statistically significant) (see Table 2 and Figure 2). Specific to Experiment 2, participants under the "calm" condition were more likely to accept offers than under the "anxious" condition. For example, combing the corresponding columns in Table 2 (Experiment 2), it reveals that the overall rejection rate in the "calm" condition (23.1%) was lower than in the "anxious" condition (30.0%) (mean difference = 6.9%, $F(1,11) \approx 4.62$, $p \approx .05$).

4 Discussion

The ultimatum game showcases the potential conflict between two of the main motives underlying social decision making: self-interest and fairness [21]. In the present study, we examined the effects of emotions in resolving such a conflict in two emotional dimensions, valence (positive vs. negative) and arousal (high vs. low), in four emotional states, "happy", "sad", "calm", and "anxious". We found that participants were more likely to accept offers in positive valence emotions ("happy" and "calm") than in negative valence emotions ("sad" and "anxious"), and the reduction of rejection rates was more apparent for "unfair" offers than "fair" offers. In addition, the reduction of rejection rates appeared to be independent of the arousal levels (high arousal in "happy" and "anxious" versus low arousal in "calm" and "sad").

In general, our findings supported our hypotheses that emotion as a separate input can causally affect decision-making, and emotional states with positive valence can alter people's attentional distribution between the fairness preference and self-interest by alleviating the perception of unfair treatment. That is, under the influence of positive valence emotions, participants were less likely to be distracted by the unfair treatment and more likely to make decisions based on their self-interest. Our results were congruent with the recent findings in both neurological and psychological research which posits that emotions can serve as a separate information input to directly shape the decision process [19, 22-25]. There is a convergence of opinions emerging from recent cognitive and affective sciences pointing toward the reciprocal causal links between the cognitive, behavioral, and somatic mechanisms, where

emotions are considered as self-perpetuating emergent systems [26], and positive affects can enhance evaluations and empower potential responses [27, 28]. Together, it is indicated that positive valence emotions, by broadening people's evaluations of decision perspectives and alleviating the perception of unfairness, may help people regain focus on their economic self-interest.

Acknowledgement. Supported by the Office of Naval Research (ONR) grant number N00014-08-1-0042, and the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021.

References

1. Camerer, C.F., Fehr, E.: When does "economic man" dominate social behavior? Science 311(5757), 47–52 (2006)
2. Güth, W., Schmittberger, R., Schwarze, B.: An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3(4), 367–388 (1982)
3. Camerer, C.F.: Behavioral game theory: Experiments in strategic interaction. Russell Sage Foundation, New York (2003)
4. Zhou, X., Wu, Y.: Sharing losses and sharing gains: Increased demand for fairness under adversity. Journal of Experimental Social Psychology 47(3), 582–588 (2011)
5. Handgraaf, M.J.J., et al.: Evaluability of outcomes in ultimatum bargaining. Organizational Behavior and Human Decision Processes 95(1), 97–106 (2004)
6. Nowak, M.A., Page, K.M., Sigmund, K.: Fairness versus reason in the Ultimatum Game. Science 289(5485), 1773–1775 (2000)
7. Rabin, M.: Incorporating fairness into game theory and economics. The American Economic Review 83(5), 1281–1302 (1993)
8. Knoch, D., et al.: A neural marker of costly punishment behavior. Psychological Science 21(3), 337–342 (2010)
9. Knoch, D., et al.: Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314(5800), 829–832 (2006)
10. Knoch, D., et al.: Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. Proceedings of the National Academy of Sciences 106(49), 20895–20899 (2009)
11. Straub, P.G., Murnighan, J.K.: An experimental investigation of ultimatum games: Information, fairness, expectations, and lowest acceptable offers. Journal of Economic Behavior & Organization 27(3), 345–364 (1995)
12. Blount, S.: When social outcomes aren't fair: The effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes 63(2), 131–144 (1995)
13. Pillutla, M.M., Murnighan, J.K.: Unfairness, anger, and spite: Emotional rejections of ultimatum offers. Organizational Behavior and Human Decision Processes 68(3), 208–224 (1996)
14. Mikula, G., Scherer, K.R., Athenstaedt, U.: The role of injustice in the elicitation of differential emotional reactions. Personality and Social Psychology Bulletin 24(7), 769–783 (1998)
15. Sanfey, A.G.: Social decision-making: Insights from game theory and neuroscience. Science 318(5850), 598–602 (2007)

16. Sanfey, A.G., et al.: The neural basis of economic decision-making in the ultimatum game. Science 300(5626), 1755–1758 (2003)
17. Wang, H., Coble, C., Bello, P.: Cognitive-affective interactions in human decision-making: A neurocomputational approach. In: Sun, R., et al. (eds.) Proceedings of the Twenty-eighth Annual Meeting of the Cognitive Science Society, pp. 2341–2346. Lawrence Erlbaum, Hillsdale (2006)
18. Fredrickson, B.L.: The broaden–and–build theory of positive emotions. Philosophical Transactions of the Royal Society of London 359(1449), 1367–1377 (2004)
19. Jefferies, L.N., et al.: Emotional valence and arousal interact in attentional control. Psychological Science 19(3), 290–295 (2008)
20. Mitterschiffthaler, M.T., et al.: A functional MRI study of happy and sad affective states induced by classical music. Human Brain Mapping 28(11), 1150–1162 (2007)
21. van Dijk, E., Vermunt, R.: Strategy and fairness in social decision making: Sometimes it pays to be powerless. Journal of Experimental Social Psychology 36(1), 1–25 (2000)
22. Bechara, A., et al.: Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 50(1-3), 7–15 (1994)
23. Weber, E.U., Johnson, E.J.: Mindful judgment and decision making. Annual Review of Psychology 60(1), 53–85 (2009)
24. Damasio, A.R.: Fundamental feelings. Nature 413(6858), 781 (2001)
25. Kringelbach, M.L.: The human orbitofrontal cortex: linking reward to hedonic experience. Nature Reviews Neuroscience 6(9), 691–702 (2005)
26. Garland, E.L., et al.: Upward spirals of positive emotions counter downward spirals of negativity: Insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. Clinical Psychology Review 30(7), 849–864 (2010)
27. Clore, G.L., Huntsinger, J.R.: How emotions inform judgment and regulate thought. Trends in Cognitive Sciences 11(9), 393–399 (2007)
28. Burgdorf, J., Panksepp, J.: The neurobiology of positive emotions. Neuroscience & Biobehavioral Reviews 30(2), 173–187 (2006)

Translation of EEG-Based Performance Prediction Models to Rapid Serial Visual Presentation Tasks

Jon Touryan, Gregory Apker, Scott Kerick, Brent Lance,
Anthony J. Ries, and Kaleb McDowell

Human Research and Engineering Directorate. U.S. Army Research Laboratory
Aberdeen Proving Ground, MD 21005, USA
{jonathan.o.touryan.ctr,gregory.apker.ctr,scott.e.kerick.civ,
brent.j.lance.civ,anthony.j.ries2.civ}@mail.mil,
Kgm8@cornell.edu

Abstract. Brain wave activity is known to correlate with decrements in behavior brought on by fatigue, boredom or low levels of alertness. Being able to predict these behavioral changes from the neural activity via electroencephalography (EEG) is an area of ongoing interest. In this study we used an established approach to predict time-on-task decrements in behavior for both a realistic driving simulator and a difficult perceptual discrimination task, utilized in many brain-computer interface applications. The goal was to quantify how well EEG-based models of behavior, developed for a driving paradigm, extend to this non-driving task. Similar to previous studies, we were able to predict time-on-task behavioral effects from the EEG power spectrum for a number of participants in both the driving and perception tasks.

Keywords: EEG, Fatigue, Power Spectral Density, Driving, RSVP.

1 Introduction

The ability to detect changes in performance induced by fatigue directly from biological markers has been an area of growing interest over recent decades. One particularly relevant application is the detection of reduced alertness or fatigue during driving. Because fatigue is a major cause of accidents and injury when operating motor vehicles [1], robust identification of fatigue before it impairs behavior would be of significant value. To this end, numerous studies have identified indicators of fatigue-induced changes in driver performance either from physiological observables [2–4] or neural signals [5]. While the physiological signals may be easier to acquire, some without necessitating direct contact with the driver, neural markers offer a more direct measure of the underlying changes in cognitive state. These changes in cognitive state, by nature, are the most proximal cause of the performance decrements and offer the best chance of detecting fatigue before the effects are evidenced in the driving behavior.

Electroencephalography (EEG) is the most common approach for quantifying the neural correlates of fatigue. Typically, EEG measurements are acquired during a long, sustained and monotonous task such as highway driving. With such tasks, behavior

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 521–530, 2013.

begins to degrade as a function of time-on-task, presumably induced by fatigue or boredom. Features of the EEG signal, such as fluctuations in power along certain frequencies or changes in evoked amplitudes, can then be correlated with the degradation in performance. Many studies exploring the neural correlates of fatigue use changes in the EEG log power spectrum as principal features in their analysis [6–9]. This idea is based on a large body of literature that has linked EEG frequency bands, such as theta (4 to 8 Hz) or alpha (8 to 13 Hz) to changes in task-relevant behavior.

One of the most general but potentially powerful approaches was described by Lin et al (2005) [10]. This approach takes an agnostic view as to the *a priori* selection of frequency bands but rather uses principal component analysis to identify the sets of frequencies that explain the most variance in the EEG power spectrum. Here, we wanted to extend the Lin approach to a non-driving, but potentially fatigue-inducing, task to quantify how well this particular EEG-based model of driver performance would extend to other domains. In this case we chose a perceptual discrimination task often utilized in brain-computer interface technology (BCIT): rapid serial visual presentation (RSVP). In RSVP, visual stimuli are presented in a rapid sequence (from 1 to 20 Hz) while the operator tries to identify the few target or relevant stimuli from the many irrelevant stimuli.

In this study, we sought to quantify how well the Lin approach for predicting time-on-task decrements in driving behavior translated to performance in an RSVP task. To accomplish this, we designed a study in which participants engaged in both a monotonous driving task and a perceptually difficult RSVP task. This construct allowed us to examine how well the Lin approach performed within and across task paradigms. Importantly, to quantify the nature and extent of the time-on-task decrements in performance, we acquired subjective, behavioral, and neurophysiologic measures throughout the experiment.

2 Methods

Participants. Eighteen participants were recruited from the general population. They ranged in age from 21 to 49 (mean = 31.1) and included seven males. Fifteen of the participants were right handed, one was left handed, and two were ambidextrous. All individuals participated in a single multi-hour session containing three phases and received compensation of $20 per hour. The voluntary, fully informed consent of the persons used in this research was obtained as required by Title 32, Part 219 of the Code of Federal Regulations and Army Regulation 70-25. The investigator has adhered to the policies for the protection of human subjects as prescribed in AR 70-25. None of the participants were excluded from the analysis due to noise, movement artifacts, or low behavioral performance.

Design and Experimental Procedures. The study design involved 3 tasks (figure 1): calibration, driving, and rapid serial visual presentation (RSVP). The calibration session was always performed first but the order of the driving and RSVP alternated for each participant.

Calibration. This task consisted of a standard driving simulator (Real Time Technologies; Dearborn, MI) with a steering wheel and foot pedal controls. In the simulator the vehicle was moving down a straight highway at a constant speed (computer controlled) in the rightmost lane. Participants were asked to maintain the vehicle position within the current cruising lane by correcting for any perturbation or drift. At random intervals a lateral perturbation was applied to the vehicle, causing it to begin to veer off course. The strength of the perturbation increased until a corrective steering adjustment was made at which point the perturbation ceased, allowing the participant to return the vehicle to the center of the rightmost lane. The perturbations would only resume once the vehicle was back in the cruising lane for at least 8 seconds. If the vehicle drifted beyond the edge of the simulated roadway, participants would receive audible feedback (i.e., rumble strip noise). The simulated environment was minimal and included no traffic or scenery in order to induce boredom and task fatigue. The calibration task consisted of a single 15 minute block and was designed to familiarize participants with the driving simulator and acquire EEG baseline activity.

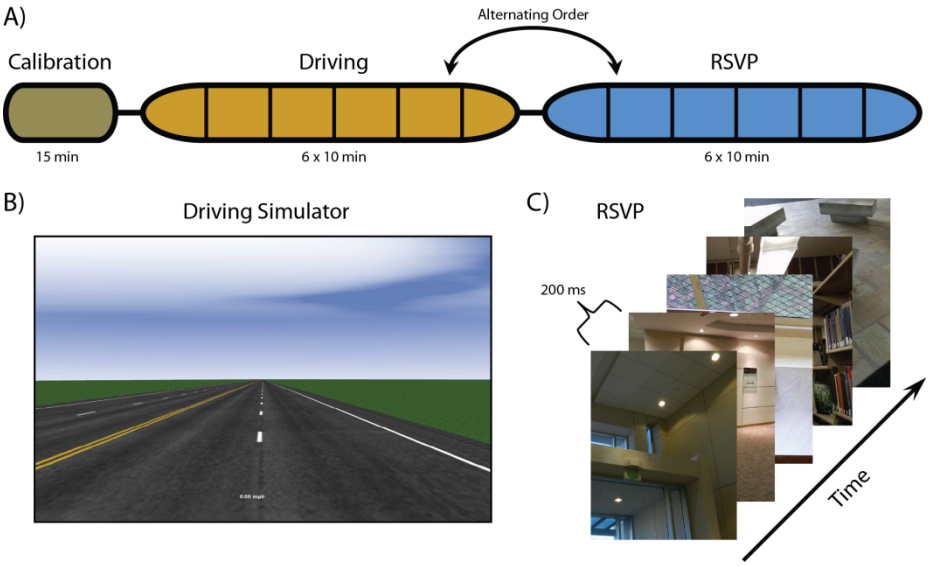


Fig. 1. Experiment design. A) Experiment timeline including calibration, driving and RSVP tasks (note that the order of the driving and RSVP tasks alternate between participants). Horizontal black lines indicate block intervals. B) Screenshot of driving simulator. C) RSVP paradigm and example images.

Driving. This task was similar to the calibration task except that participants were now given control over the vehicle speed (indicated by a speedometer). Participants were asked to maintain both the vehicle position and speed. Speed limit signs were posted at regular intervals with values of either 25 or 45 miles per hour. Again, the simulated environment was minimal and included no traffic or scenery. The driving task consisted of 6 blocks of 10 minutes each.

RSVP. This task consisted of a rapid presentation of color photographs (512 x 662 pixels) of indoor and outdoor scenes. The images were presented at 5 Hz (200 milliseconds per image) and subtended a visual angle of approximately 9°. Every 10 seconds a blank screen with the word "blink" was presented to give participants a chance to blink without missing stimuli. The RSVP task consisted of 6 blocks of 10 minutes each (to mirror the driving task). All scenes contained only inanimate objects and were manually centered, scaled and cropped. Some scenes contained target objects and others did not. Before each block participants were instructed as to the class of target objects for that block. The target classes for this experiment were: stair, container, poster, chair, and door. Before the task began, participants were familiarized with exemplars from each target class. During the RSVP, participants were instructed to press a button only when they saw an object from the current target class. The order of the target classes was randomly chosen for each participant (blocks 1-5); however, the last block (block six) always had the same target class as the first block. In addition to target class, target probability varied across each block. Six target probability values (0.01, 0.03, 0.05, 0.07, 0.09 and 0.11), one for each block, were randomly assigned at the beginning of the task.

Subjective Measures. Self-reports of fatigue (the Visual Analog Scale for Fatigue [11], the Task-Induced Fatigue Scale [12], and the Karolinska Sleepiness Scale [13]) were collected at various times during the experiment. Overall, the participant reports of fatigue confirmed that both the driving and RSVP task induced fatigue and boredom (data not shown).

Behavioral Measures. During the driving simulator task, various vehicle status measurements were acquired at 100 Hz. Since the objective was to maintain vehicle position within the rightmost lane, lane deviation (the difference between the vehicle's lateral position and the center of the lane) was the metric used to assess driver performance. During the RSVP task, participants pressed a button only when they saw a target object. Accuracy, reaction time (RT) and duration were determined from this button response. Because the image duration (200 ms) was much less than the average RT, button responses were assigned to images in the following manner. For each button press, images within the time window of 1000 ms to 300 ms preceding the response were identified. If one or more of these preceding images was a target, the button press was assigned to the first (oldest) target image. RT was then calculated from the onset of that target image. If no targets occurred within the preceding time window, the button press was assigned to the non-target image that preceded the button press by 600 ms. However, due to the ambiguity of assigning a button press to a non-target image, no RT's were calculated for these images and the process was only used to determine the false alarm rate.

Electroencephalography Measures. Electrophysiological recordings were digitally sampled at 1024 Hz from 256 scalp electrodes over the entire cortex using a BioSemi Active Two system (Amsterdam, Netherlands). External leads were placed on the

outer canthus, above and below the orbital fossa of the right eye to record electrooculography (EOG). For power spectrum analysis, EEG was referenced to the average mastoids and down-sampled to 256 Hz using the EEGLAB toolbox [14]. For power spectrum analysis, only data from the midline channels A1, A6 and A21 (roughly corresponding to Cz, Pz, and Oz) was utilized. This selection of channels was based on previous studies that examined the correlation between EEG power and driver performance over the scalp [10].

Moving-average power spectra and linear regression models were calculated in the same way as described by Lin et al (2005) with two minor exceptions. First, to normalize the power spectral density (PSD) estimates, we calculated the z-score at each frequency over all PSD epochs [15]. Second, for improved signal-to-noise ratio, only frequencies between 5-50 Hz from the three channels were included in the principal component analysis (PCA). Each participant had separate models fit to their PSD and behavioral data from the driving and RSVP tasks. For the driving task, the model used absolute lane deviation as the behavioral metric. For the RSVP task, separate models were built using target accuracy, reaction time and duration of button press.

To reduce potential overfitting of the 50-order linear model, the EEG and behavioral data was split into 6 cross-validation sets corresponding to the experimental blocks. Specifically, a model would be built with data from 5 blocks and tested on data from the remaining block. The model performance was quantified using Pearson's correlation coefficient. The model with the highest correlation coefficient between the predicted and actual behavior was selected for subsequent cross task prediction.

3 Results

3.1 Self Reports and Behavior

To capture behavioral performance fluctuations induced by changes in alertness level, we averaged the RSVP behavioral metrics within sliding 90 second windows (using 45 second step size). In addition to being aligned with previous studies [8, 10] this window size enabled a robust estimate of accuracy (average number of trials per window = 445) even when the target probability was low. Because participants were only required to respond when they saw a target and the attribution process for false alarms was ambiguous (see Methods), behavioral metrics were calculated for correct target trials. Notably, there were large fluctuations both within and across blocks (data not shown). Across blocks, these fluctuations could be due to changes in either task parameters (target type or target probability) or alertness level (due to fatigue or boredom). However, within block fluctuations must be precipitated from endogenous changes such as perceptual learning, fatigue, or boredom. To isolate these endogenous changes, linear fits and corresponding significance levels were calculated for each block.

Over the population, the clear modulator of performance in the RSVP task was target class. This was true for accuracy ($F(1,17) = 22.33$, $p < 0.001$), RT ($F(1,17) = 7.41$, $p < 0.001$) and to a lesser extent button duration ($F(1,17) = 2.67$, $p < 0.05$). Similarly, target probability had a significant effect on RT ($F(1,17) = 25.11$, $p < 0.001$) and

duration ($F(1,17) = 45.27$, $p < 0.001$) but not accuracy ($F(1,17) = 1.48$, $p = 0.23$). Although none of the behavioral metrics showed a significant time-on-task effect across blocks, most participants had at least one block with a significant decrease in accuracy or increase in RT, reflecting a time-on-task performance decrement (average number of blocks per participant = 1.94). Far fewer participants had blocks with significant performance improvements, either through increased accuracy or decreased RT (average number of blocks per participant = 0.56). This difference was significant ($p < 0.01$; paired t-test,), indicating that fatigue or boredom rather than perceptual learning modulated within block performance variations.

3.2 PSD Regression Model

Previous studies have shown a strong relationship between the EEG power spectrum and time-on-task decrements in performance, especially in monotonous driving tasks [16]. Less is known about the link between the EEG power spectrum and behavior in more complex perceptual tasks, such as the RSVP paradigm described here. To explore this relationship further, we fit a linear model to continuous estimates of the EEG power spectral density (PSD) from the driving task (see Methods). This model was similar to ones previously reported in the literature and utilized PSD estimates from three midline electrodes (Cz, Pz, and Oz) to predict the absolute value of the lane deviation (a standard metric of driver performance). To match RSVP behavioral metrics described above, both the PSD estimates and lane deviation were smoothed over a similar 90 second window. Here, individual models were built for each participant using a 6 fold cross-validation scheme (see Methods). The 6 models were used to predict lane deviation data not included in the training set, yielding 6 independent prediction scores (correlation coefficients). Notably, there was a large variance of prediction performance within the set of models for each participant. This indicated that the underlying relationship between the PSD and driving performance was highly variable between training and testing sets and/or the linear model tended to over-fit the training data. Since the purpose of this study was to explore how well general models that link EEG power spectra with behavior translate across tasks, we did not take additional measures to isolate which model coefficients added significant predictive power (e.g., step-wise regression). Rather, to minimize over-fitting for each participant we chose the model with the highest prediction score on the validation dataset. This model was then used to predict the participant's behavior in the RSVP task.

In parallel, we used the same approach to build predictive models of behavior for the RSVP task. These PSD models had the same structure and cross-validation scheme as described above. As with the PSD models of driving behavior, these models show a high degree of variability for each participant. Again, to minimize overfitting, we chose the model with the highest prediction score on the validation dataset. Over the population there was no significant difference in model performance between the two tasks (driving r = 0.547 ± 0.139, RSVP r = 0.556 ± 0.236, $p = 0.74$; Wilcoxon signed rank test). There was also no significant difference between RSVP models trained on accuracy versus RT (accuracy r = 0.556 ± 0.236, RT r = 0.504 ± 0.189, $p = 0.45$) or accuracy versus duration (accuracy r = 0.556 ± 0.236, duration

r = 0.569 ± 0.250, p = 0.71). The results indicate that these linear models of behavior based on the EEG power spectrum have roughly the same predicative power in both tasks and across a number of metrics.

To more directly test how these models generalize across tasks, we used the models from the driving task to predict RSVP behavior and models from the RSVP task to predict driving behavior. To accomplish this we took the best model for each participant within a given task (driving or RSVP) and applied it to the PSD estimate from the alternate task. Importantly, we needed to adapt the behavior prediction estimate to the new type of behavior. For driving, the metric (absolute lane deviation) typically increases with time-on-task fatigue or boredom; alternatively, the RSVP metric (target accuracy) typically decreases under the same conditions. Thus, we added an additional affine fit of the model prediction to match the novel behavior metric. Figure 2 shows an example of a model built on driving data used to predict RSVP behavior. Here, the driving model predicted the RSVP behavior with a reasonable degree of accuracy, capturing the fluctuation in behavior both within and across blocks. While the behavior and PSD estimates integrate data from a 90 second window, this type of linear model could have predictive power that either leads or lags the behavior. To assess the causality of the prediction relative to the actual behavior, we performed a cross-correlation analysis.

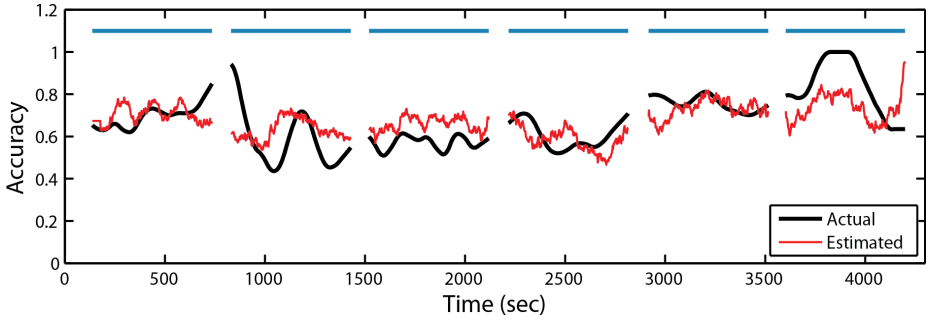


Fig. 2. Cross-task prediction for one participant. Actual (black line) and estimated (red line) target accuracy for the RSVP task. Estimated behavior was derived from model fit to driving data for the same participant (see Methods). Horizontal bars indicate experiment blocks.

Over the population, the PSD models derived from driving task were able to predict some behavioral variance in the RSVP task and vise versa (figure 3). To establish significance, a bootstrap reshuffling technique was applied to the predicted behavior and smoothed by the same integration window as the PSD. In this way, a distribution of random correlation coefficients could be parameterized and a confidence interval established for each prediction. For the driving task models, the average correlation coefficient between the prediction and RSVP behavior was 0.205 ± 0.189. Only 2 of the 18 models had a significant correlation coefficient ($p < 0.01$). For the RSVP task models, the average correlation coefficient between the prediction and the driving behavior was 0.211 ± 0.164. Here, 5 of the 18 models had a significant correlation coefficient. By using the optimal temporal lag (peak of the cross-correlation curve)

prediction scores improved slightly (driving model r = 0.282 ± 0.146, RSVP model r = 0.332 ± 0.166). The optimal temporal lags for both the driving and RSVP-based models were within the 90 second integration window (driving model lag = 70.2 seconds, RSVP model lag = 58.1 seconds). Here, positive values indicate that the predicted behavior leads to that actual behavior.

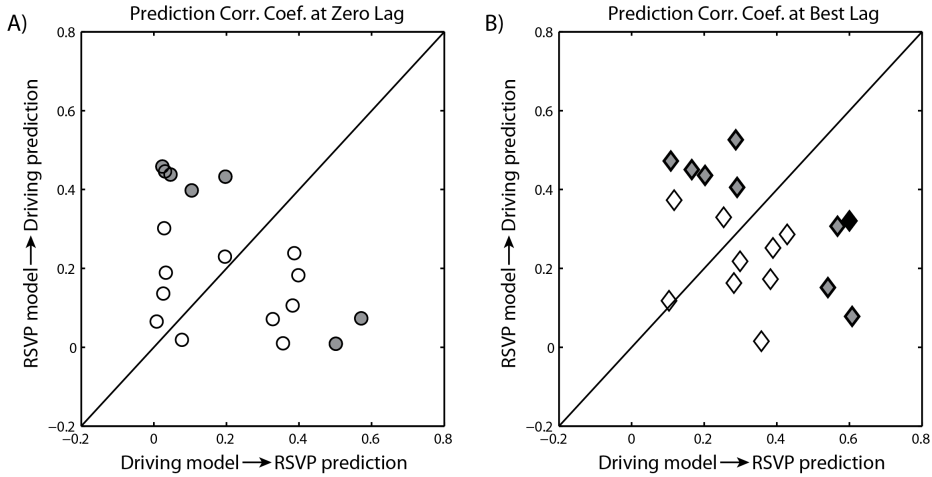


Fig. 3. Cross-task prediction over the population. A) Correlation coefficients of the cross-task predictions for each participant with no temporal shift. B) Correlation coefficients of the cross-task predictions for each participant at best temporal alignment. Level of shading indicates significance: black = significant for both conditions, gray = significant for one condition, clear = not significant.

4 Discussion

Using an established approach based on the EEG log power spectrum, we were able to predict fluctuations in driver performance with a reasonable degree of accuracy (correlation coefficients greater than 0.5) for a number of participants. Likewise, we were able to predict fluctuations in target accuracy in the RSVP task to a similar degree. Furthermore, when models fit under the driving paradigm were applied to the RSVP task, predictive power remained significant for some participants, despite its reduction overall (figure 3).

Interestingly, there was a large variation in model performance during the fitting and cross-validation processes. This may have resulted from several factors. First, the validation sets from which the scores were derived were relatively short (approximately 10 min long). If there was a significant change in behavioral dynamics (e.g., less or more time-on-task fatigue) between the training and testing sets, the predictions could differ substantially. Second, the linear regression model described here always included 50 components (top 50 eigenvectors of the PSD). In some instances this feature space may have been over-represented leading to a number of coefficients lacking significant predictive value. For this reason, some approaches use step-wise

regression [17] to minimize over-fitting. Third, while the PSD estimation process utilized a median filter and the power spectra were smoothed over a 90 second integration window, noise and artifacts could have degraded the model fitting process. Independent component analysis (ICA) has likewise been used in this paradigm to mitigate artifacts and improve predictions [18].

It is important to note that the major source of variance in the RSVP paradigm was not time-on-task. Instead, target class was the strongest modulator of behavioral variance across blocks. While the target stimuli were roughly matched along low-level visual dimensions such as luminance, object size and eccentricity, we observed a significant performance difference in all three metrics (accuracy, RT and button duration). Likewise, the changes in target frequency across blocks significantly modulated behavior. Thus, given such exogenous sources of behavioral variance, it is encouraging that the Lin et al (2005) approach maintains significant predictive power in this more complex task.

The Lin approach we used employs a relatively simple model for predicting behavior (i.e., a linear regression of the PSD coefficients along a single behavioral metric). Since that original study, this method has been extended to incorporate ICA [18] and fuzzy neural-networks [19]. However, the linear approach still represents a solid and interpretable framework to explore the relationship between the EEG power spectra and behavior in a variety of tasks. In addition, this method is computationally simple and utilizes universal signal processing components (such as PSD estimation). Thus, as we have demonstrated here, it remains a practical approach for an embedded application in a real-time EEG processing system [20].

References

1. Connor, J.: Driver sleepiness and risk of serious injury to car occupants: population based case control study. BMJ 324, 1125–1125 (2002)
2. Sommer, D., Golz, M.: Evaluation of PERCLOS based current fatigue monitoring technologies. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4456–4459 (2010)
3. Vogel, A.P., Fletcher, J., Maruff, P.: Acoustic analysis of the effects of sustained wakefulness on speech. The Journal of the Acoustical Society of America 128, 3747–3756 (2010)
4. Vural, E., Çetin, M., Erçil, A., Littlewort, G., Bartlett, M., Movellan, J.: Machine Learning Systems for Detecting Driver Drowsiness. In: Takeda, K., Erdogan, H., Hansen, J.H.L., Abut, H. (eds.) In-Vehicle Corpus and Signal Processing for Driver Behavior, pp. 97–110. Springer, US (2009)
5. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neuroscience & Biobehavioral Reviews (in Press)
6. Balasubramanian, V., Adalarasu, K., Gupta, A.: EEG based analysis of cognitive fatigue during simulated driving. International Journal of Industrial and Systems Engineering 7, 135–149 (2011)
7. Jap, B.T., Lal, S., Fischer, P., Bekiaris, E.: Using EEG spectral components to assess algorithms for detecting fatigue. Expert Systems with Applications 36, 2352–2359 (2009)

8. Jung, T.-P., Makeig, S., Stensmo, M., Sejnowski, T.J.: Estimating alertness from the EEG power spectrum. IEEE Transactions on Biomedical Engineering 44, 60–69 (1997)
9. Lal, S.K.L., Craig, A.: Reproducibility of the spectral components of the electroencephalogram during driver fatigue. International Journal of Psychophysiology 55, 137–143 (2005)
10. Lin, C.T., Wu, R.C., Jung, T.P., Liang, S.F., Huang, T.Y.: Estimating driving performance based on EEG spectrum analysis. EURASIP Journal on Advances in Signal Processing 2005, 3165–3174 (2005)
11. Monk, T.H.: A visual analogue scale technique to measure global vigor and affect. Psychiatry Research 27, 89–99 (1989)
12. Matthews, G., Desmond, P.A.: Personality and multiple dimensions of task-induced fatigue: a study of simulated driving. Personality and Individual Differences 25, 443–458 (1998)
13. Akerstedt, T., Gillberg, M.: Subjective and Objective Sleepiness in the Active Individual. International Journal of Neuroscience 52, 29–37 (1990)
14. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134, 9–21 (2004)
15. Peiris, M.T.R., Davidson, P.R., Bones, P.J., Jones, R.D.: Detection of lapses in responsiveness from the EEG. Journal of Neural Engineering 8, 016003 (2011)
16. Ting, P.-H., Hwang, J.-R., Doong, J.-L., Jeng, M.-C.: Driver fatigue and highway driving: A simulator study. Physiology & Behavior 94, 448–453 (2008)
17. Stikic, M., Johnson, R.R., Levendowski, D.J., Popovic, D.P., Olmstead, R.E., Berka, C.: EEG-Derived Estimators of Present and Future Cognitive Performance. Front Hum. Neurosci. 5 (2011)
18. Lin, C.T., Wu, R.C., Liang, S.F., Chao, W.H., Chen, Y.J., Jung, T.P.: EEG-based drowsiness estimation for safety driving using independent component analysis. IEEE Transactions on Circuits and Systems I: Regular Papers 52, 2726–2738 (2005)
19. Lin, C.-T., Ko, L.-W., Chung, I.-F., Huang, T.-Y., Chen, Y.-C., Jung, T.-P., Liang, S.-F.: Adaptive EEG-Based Alertness Estimation System by Using ICA-Based Fuzzy Neural Networks. IEEE Transactions on Circuits and Systems I: Regular Papers 53, 2469–2476 (2006)
20. Lin, C.-T., Chang, C.-J., Lin, B.-S., Hung, S.-H., Chao, C.-F., Wang, I.-J.: A Real-Time Wireless Brain-Computer Interface System for Drowsiness Detection. IEEE Transactions on Biomedical Circuits and Systems 4, 214–222 (2010)

Adult Neurogenesis: Implications on Human And Computational Decision Making

Craig M. Vineyard[1], Stephen J. Verzi[1], Thomas P. Caudell[2], Michael L. Bernard[1], and James B. Aimone[1]

[1] Sandia National Laboratories
Cognitive Modelling Department
Albuquerque, New Mexico, USA
cmviney@sandia.gov
[2] Electrical and Computer Engineering Department
University of New Mexico
Albuquerque, New Mexico, USA

Abstract. Adult neurogenesis is the incorporation of new neurons into established, functioning neural circuits. Current theoretical work in the neurogenesis field has suggested that new neurons are of greatest importance in the encoding of new memories, particularly in the ability to fully capture features which are entirely novel or being experienced in a unique way. We present two models of neurogenesis (a spiking, biologically realistic model as well as a basic growing feedforward model) to investigate possible functional implications. We use an information theoretic computational complexity measure to quantitatively analyze the information content encoded with and without neurogenesis in our spiking model. And neural encoding capacity (as a function of neuron maturation) is examined in our simple feedforward network. Finally, we discuss potential functional implications for neurogenesis in high risk environments.

Keywords: Neurogenesis, Dentate Gyrus, Information Theoretic Complexity, Neural Network Modeling.

1 Introduction

Human cognition is facilitated by numerous forms of neuronal plasticity that span many different scales in both spatial and temporal dimensions. One such neural process that has received considerable attention over the past decade is adult neurogenesis, which is the incorporation of new neurons into established, functioning neural circuits [1]. Neurogenesis is uniquely limited to a few regions and has been shown to be regulated by a wide range of intrinsic and extrinsic behavioral conditions. The most studied neurogenic region is the dentate gyrus (DG) area of the hippocampus, a region known to be critically important for learning and memory.

Current theoretical work in the neurogenesis field has suggested that new neurons are of greatest importance in the encoding of new memories, particularly

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 531–540, 2013.

in the ability to fully capture features which are entirely novel or being experienced in a unique way. This type of memory has the potential to be of critical importance in high consequence scenarios, in particular in situations where the decision-maker must base their reasoning on novel, previously unexperienced, aspects of the environment. Without the capability to properly encode and process novel components of an experience, a decision-maker may fall back on the familiar, which while often a proper strategy, can sometimes be detrimental.

We anticipate that this work can impact cognitive function in two distinct respects: 1) interventions that increase neurogenesis promise to be an effective method to improve acute decision making by individuals; and 2) computational approaches that implement neurogenesis-like plasticity and structural dynamics can potentially motivate a powerful new form of algorithms that can facilitate data processing and decision-making assistance in revolutionary ways.

In the following sections we will briefly describe the neurophysiology involved in neurogenesis and an associated encoding hypothesis, provide a real world high consequence decision making scenario with potential functional implications, describe two computational models to investigate neurogenesis, and provide some results analyzing these models.

2 Neurophysiology

Situated within the medial temporal lobe, the hippocampus is a well studied neural region that receives an amalgamation of sensory input signals and contributes significant functional importance such as its key role in episodic memory formation [2]. The dentate gyrus (DG) serves as an entry region of the hippocampus receiving sensory stimuli from both lateral and medial entorhinal cortex (EC) [3]. Although it is comprised of several cell types, granule cells are the most populous neuron types within the DG. The DG is a relatively large region (consisting of approximately 10 to 20 million neurons in humans), however it exhibits sparse activation meaning that only approximately 2 percent of these neurons are active at a given instance [4]. DG activity subsequently serves as input to the highly recurrent cornu ammonis 3 (CA3) region of hippocampus for further processing.

The sparse activation of the DG has often been attributed to a pattern separation functionality within DG [5]. From this perspective, the relatively few neurons firing despite the large size of DG corresponds to a unique non-overlapping encoding of the multi-modal sensory inputs from EC. Alternatively, another proposed role for DG is to control memory resolution [6]. From this perspective, young immature neurons are hyperexcitable and broadly respond to a wide variety of input stimulation. Mature neurons, on the other hand, are narrowly tuned to respond to specific inputs they have learned to selectively fire to. The integration of both mature and young neurons within the same neural network allows for a mixed coding hypothesis. From this perspective, the young, easily excitable neurons are integral for incorporating new memories within a neural network without interfering with existing encodings represented by the mature neurons.

3 Real World Scenario

As a real world example of a high consequence decision-making scenario with potential implications for neurogenesis, consider the role of the drone operator. Rather than piloting their aircraft internally from the confines of a cockpit, as conventional pilots do, drone pilots remotely operate their aircraft from a distant workstation with real time video feeds projected on computer screens. Some drones are equipped with weapons and are consequently able to take action if a hostile target is detected.

However, the majority of a drone operator's time is spent watching and surveying. According to Massachusetts Institute of Technology (MIT) aeronautics and astronautics professor Mary Cummings, "You might park a UAV over a house, waiting for someone to come in or come out, and thats where the boredom comes in" [7]. Despite the similar environment to that of a video game, a drone operator's shift is typically less action intensive. Instead, "...it is not uncommon in search and reconnaissance missions for a UAV pilot to spend the majority of the mission waiting for a system anomaly to occur, with only occasional system interactions" [8].

It is of crucial importance that this rare, anomalous event consisting of a target of interest appearing in what is an otherwise highly familiar environment does not go undetected. It is possible that highly active young neurons may facilitate the ability to encode and perceive this novel, but significant event.

4 Computational Models

To investigate the possible functional significance of neurogenesis we have developed two neural network models. The first is a large scale, biologically realistic, spiking neural network model. The second is a simple rate-coded feedforward network that grows new neurons and connections. In the following we will describe the two networks in greater detail.

4.1 Spiking Dynamics Model

We have developed a biologically motivated, spiking model comprised of nine cell types representing EC inputs as well the molecular layer, granule cell layer, and hilus of DG. The underlying neuron model we have implemented uses Izhikevich neural dynamics so that we can fit to actual electrophysiology data from mature and immature granule cells and hilar interneurons [9]. We have also incorporated biologically realistic ratios of neurons within the model. Particularly, we have experimented with a model consisting of 5,500 EC neurons and 50,000 DG granule cells. The EC neurons are split between lateral and medial EC providing object cell and grid cell inputs respectively.

The particular input firings are driven by a multi-context multi-day simulated experimental paradigm. In a single simulation day, the model is presented three different contexts consisting of a variety of items in various locations. Over the

course of multiple simulation days, the first context is the same every day and is repeatedly presented to the model as a very familiar input. The second input presented each day is familiar context that the model has been presented before, but not as frequently as the first very familiar input. And finally, the third input presented each day is a novel, formerly unseen input (although it may consist of some formerly seen items in new locations and paired with different combinations of items). This experimental paradigm allows the model to investigate both acute and long term effects of neurogenesis while varying neurogenesis rates in a controlled manner.

4.2 Basic Neurogenesis Model

Additionally, to investigate fundamental neurogenesis functionality, such as the mixed coding hypothesis, we have also implemented a basic feedforward model which relaxes biological realism. This simplistic model consists of two layers of neurons. A fixed size input layer representing the EC, and a growing layer of DG granule cells.

The EC layer consists of both excitatory and inhibitory inputs, with four times as many excitatory as inhibitory inputs. Both the excitatory and inhibitory neurons exhibit a twenty percent activation each timestep. Over time, the DG layer grows both by incorporating new neurons as well as adding additional synapses (both excitatory and inhibitory) to the existing neurons as they mature. Just as there are more excitatory inputs than inhibitory, there are likewise more synapses to excitatory inputs than inhibitory. However, the inhibitory synapses have a stronger effect than the excitatory synapses. Throughout the neurogenesis network growth process, these ratios are preserved.

As a simplistic model of neural behavior, DG neurons fire if for a given timestep, input excitation exceeds input inhibition. This basic behavior is subsequently regulated by Hebbian learning such that if an input causes a DG neuron to fire, its synapses are updated accordingly.

Over time, new neurons are added to the DG layer. Each of these new neurons is randomly connected to the EC inputs, with a baseline amount of synapses. Additionally, over time, all neurons incorporate new synapses until they reach full maturity which happens when synaptic connections to the EC inputs reach twenty percent (of all available EC inputs). Throughout this temporal maturation process, a set of EC inputs are cycled through. Rather than exposing all of the inputs to the full input set, instead subsets of the inputs are presented during certain time windows.

For a more detailed description of this model please see [10].

5 Results

To quantitatively assess the potential benefits to neurogenesis we have analyzed computational complexity as an estimation of information representation from

an information theoretic standpoint as well as examined neuronal encoding rates. The results of these analyses for our two computational models are presented next.

5.1 Computational Complexity of the Spiking Dynamics Model

To analyze the encoding capability of our spiking dynamics model we looked at the computational complexity of the granule cell neural ensemble over the course of the presentation of a particular context. Shannon entropy is a fundamental approach to quantize the amount of information in a variety of sources such as communication channels [11]. Additionally, many approaches have been devised to apply this sort of information measure to neurons [12]. However, doing so requires knowledge of the firing behavior probability distribution for the neurons within the model.

Rather, in lieu of estimating neuron firing probabilities, we have used complexity as a measure of compressibility in order to estimate entropy to quantitatively assess the information content of a signal. Szczepanski et al. applied the general Lempel-Ziv complexity (LZ-Complexity) measure to estimate entropy of real and simulated neurons [13]. But unlike the work of Szczepanski et al., rather than applying LZ-Complexity analysis to individual neuron spike trains, we have applied the approach to a neural population as a whole. LZ-Complexity is based upon measuring the rate of generation of new patterns along a sequence of characters in a string being compressed [14]. Applied to neuron spike trains, this technique looks for repeated spiking behavior over time. Instead, by applying it across an entire neural ensemble, we assessed repeated patterns of neural co-activity. Synaptic modifications alter the firing behavior of the neural network through learning. In order to account for this plasticity of the network, rather than computing the ensemble complexity at each timestep, we concatenated all of the firing outputs of the entire neural ensemble (while presented a single input context) into a long spike signal. This approach is depicted in Fig. 1.

Once the spike signal is converted into a binary signal, where an action potential is encoded as a one and the absence of activity by a zero, the normalized complexity may then be computed as follows:

$$c_\alpha(x^n) = \frac{C_\alpha(x^n)}{n} * log_\alpha n. \tag{1}$$

Normalized complexity measures the generation rate of new patterns along a word of length n with letters from an alphabet of size α (in this case two). Additionally, it can be proven [11] that as the string length (our series of neural firings in this case) goes to infinity, the supremum of the normalized complexity approaches the entropy of the signal S:

$$\limsup_{n\to\infty} c_\alpha(x^n) \leq H_\alpha(S). \tag{2}$$

We have implemented two instantiations of a biologically inspired spiking neural model, each consisting of 50,000 granule cells. The difference between these two

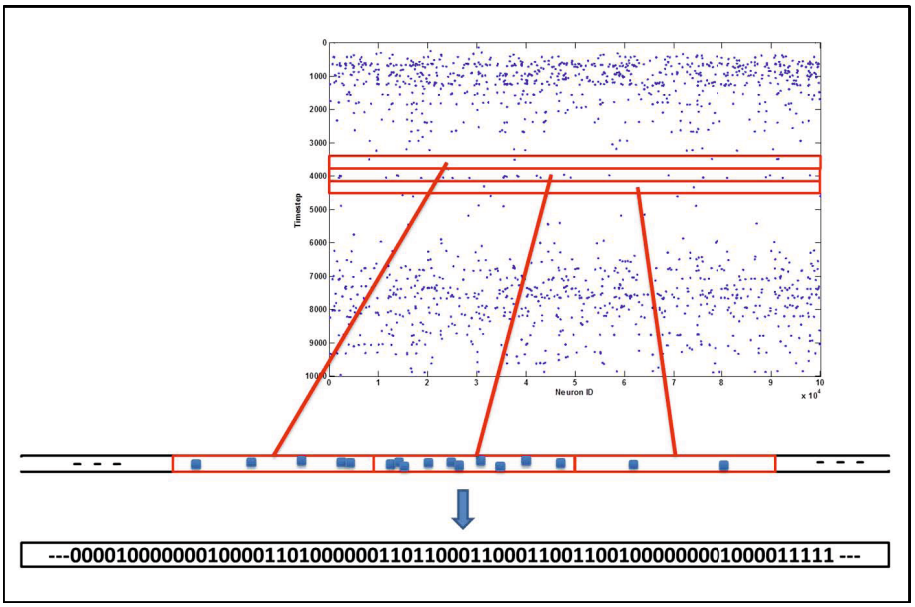


Fig. 1. Concatenation of neural firings across the population ensemble to generate a binary spike signal preserving temporal synchrony.

models is that the first does not implement neurogenesis while the second has a ten percent neurogenesis rate. Both models were exposed to three contexts across different simulated days as described formerly in the model description. Fig. 2 depicts the normalized complexity values for these two models across seven days of contexts (with the three numbers corresponding to normalized complexity for each of the three contexts, respectively). As evident by Fig. 2, the neural network with neurogenesis exhibits a distinct increase in information content, quantitatively inferred by means of normalized complexity, compared to the network with no neurogenesis across all days and contexts.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
No NG	0.000003 0.000003 0.000003	0.000002 0.000002 0.000003	0.000002 0.000002 0.000002	0.000002 0.000002 0.000003	0.000002 0.000002 0.000002	0.000002 0.000003 0.000002	0.000002 0.000002 0.000002
10% NG	0.000032 0.000031 0.000031	0.000033 0.000035 0.000033	0.000861 0.000618 0.000747	0.000898 0.000846 0.000748	0.000789 0.000872 0.000824	0.001201 0.000458 0.000676	0.001298 0.001250 0.001050

Fig. 2. Normalized Complexity values for 50,000 granule cell network with zero and ten percent neurogenesis over seven days of varied contexts

5.2 Basic Neurogenesis Model

In evaluating our basic neurogenesis model, we experimented with an EC size of 12,500 neurons (10,000 excitatory and 2,500 inhibitory inputs). The EC layer itself had no input, however the patterns of activity we specified for it at a given timestep served as the inputs for the DG layer. The EC does not receive direct sensory input, but rather receives signals which have been pre-processed, such as by the visual cortex. Alternatively, specific input patterns such as visual images could be applied as inputs to the model if an appropriate neural sensory processing function (such as a hashing function) were used to process the raw input. Such a framework is illustrated in Fig. 3, where we have currently only implemented the portion to the right of the human comprehensible images in the figure, with binary EC activation patterns and a growing network of DG neurons.

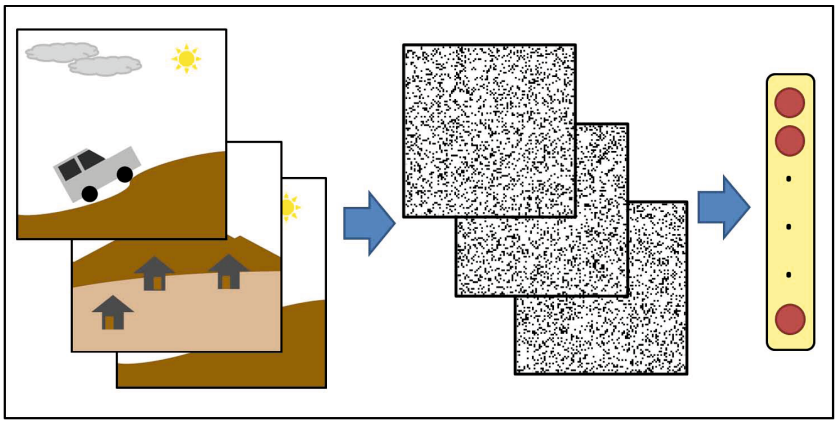


Fig. 3. General framework of a basic neurogenesis model proceeding from input images to an entorhinal cortex distributed representation and subsequent processing by a growing dentate gyrus network

In our analysis, we varied the maximum allowed growth in the DG layer to investigate network plasticity as well as learn-ability. In all cases, the younger immature neurons proved to be more excitable and responsive to a greater number of inputs despite having fewer connections than the more mature neurons. And likewise, the older neurons, through maturation, became narrowly tuned and responsive to specific input stimuli. This behavior is evident in Fig. 4 where the horizontal axis delineates the particular neurons by maturation age and the vertical axis represents the number of inputs each neuron responds to. The neurons in the figure are ordered by when they were added to the network, so the older (more mature) neurons which were added first are on the left and the younger more recently added neurons are on the right. To account for variability

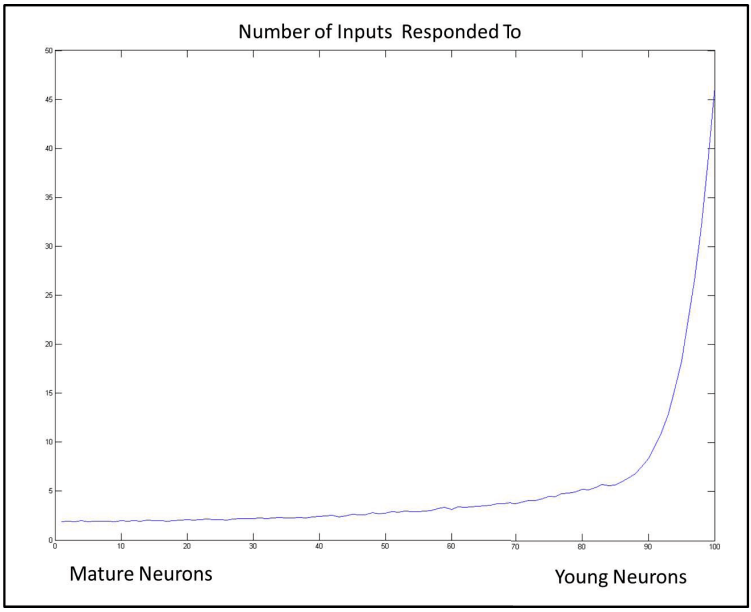


Fig. 4. Number of input patterns each neuron responds to. Moving from left to right represents the ordering in which new neurons were added to the network such that the rightmost neurons are the youngest neurons.

in the random synaptic connectivity, the number of inputs the neurons respond to are averaged over 1000 simulation runs of the model.

Furthermore, informal evaluations have also shown that given a sufficient neurogenesis rate in conjunction with an adequately sized DG proved to be sufficient to encode all inputs. This characteristic is important for network stability such that as neurons within the network mature and become tightly tuned to specific inputs that prior information is not lost in exchange for the novel stimuli. In this sense, the mature neurons are selectively responsive to narrowly tuned inputs but do not respond to novel stimuli. We evaluated this functionality by turning off Hebbian learning and re-showing the network the formerly seen inputs as well as a set of novel inputs. The mature neurons only responded to their select inputs while the younger neurons were responsive to novel stimuli as well as the formerly seen inputs.

6 Conclusions

Through neurogenesis, it appears that the incorporation of new cells within a neural circuit may be a means to increase the information content of the network as well as provide a means to encode novel stimuli. New neurons, which are highly excitable, have an increased likelihood of encoding current stimuli.

Consequently, as they mature they become more tightly tuned to particular inputs being learned and are not as easily able to incorporate the novel stimuli into the network without neurogenesis. Such a phenomena may play a crucial role in high consequence decision making scenarios such as that of a drone operator. On a surveillance mission, the majority of the images a drone operator sees may be routine and familiar if they have surveyed the same area previously. The scenario may be entirely familiar if nothing has changed regarding the area under consideration. Or it may be a highly familiar scene in which all of the usual components are there but a suspect's vehicle is parked on the other side of the house for example. However, it is of utmost importance in this domain to be cognizant of the subtle change in which a key target appears in what was an otherwise routine surveillance so the situation can be properly assessed and the appropriate action taken. A better understanding of the neurogenesis phenomena and its functional implications may allow for this capability to be increased, or alternatively incorporated within computational tools as an aid leading to improved human performance.

While the potential benefits adult neurogenesis may provide are exciting, its functional implications are still far from understood. As future work, we plan to investigate the effects of varying neurogensis rates, examine whether the maturation rate has any effect on learnability or stability, study strategies for synaptic growth/formation, and consider neuronal death as a contrasting balance to neurogenesis. Additionally, we also plan to investigate possible application areas which may benefit from neurogensis like mechanisms such as memory management, computational encoding schemes, and dynamic decision making.

Acknowledgements. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. Aimone, J.B., Jessberger, S., Gage, F.H.: Adult Neurogenesis. Scholarpedia 2(2) (2100)
2. Eichenbaum, H.: The Hippocampus and Declarative Memory: Cognitive Mechanisms and Neural Codes. Behavioural Brain Research, 199–207 (2001)
3. Suzuki, W., Eichenbaum, H.: The Neurophysiology of Memory. Annals of the NY Academy of Sciences, 175–191 (2000)
4. Andersen, P., Morris, R., Amaral, D., Bliss, T., O'Keefe, J. (eds.): The Hippocampus Book. Oxford University Press, USA (2006)
5. O'Reilly, R.C., McClelland, J.L.: Hippocampal Conjunctive Encoding, Storage, and Recall: Avoiding a Trade-Off. Hippocampus 4(6), 661–682 (1994)
6. Aimone, J.B., Deng, W., Gage, F.H.: Resolving New Memories: A Critical Look at the Dentate Gyrus, Adult Neurogenesis, and Pattern Separation. Neuron. 70(4), 589–596 (2011)

7. MIT news: Driving drones can be a drag, http://web.mit.edu/newsoffice/2012/boredom-and-unmanned-aerial-vehicles-1114.html
8. Cummings, M.L., Mastracchio, C., Thornburg, K.M., Mkrtchyan, A.: Boredome and Distraction in Multiple Unmanned Vehicle Supervisory Control. Interacting with Computers 25(1), 34–47 (2013)
9. Izhikevich, E.M.: Simple Model of Spiking Neurons. IEEE Transactions on Neural Networks 14(6), 1569–1572 (2003)
10. Li, Y., Aimone, J.B., Xu, X., Callaway, E.M., Gage, F.H.: Development of GABAergic Inputs Controls the Contribution of Maturing Neurons to the Adult Hippocampal Network. Proceedings of the National Academy of Sciences 109(11), 4290–4295 (2012)
11. Cover, T., Thomas, J.: Elements of Information Theory. Wiley, New Jersey (2006)
12. Victor, J.D.: Approaches to Information-Theoretic Analysis of Neural Activity. Biological Theory 1(3), 302–316 (2006)
13. Szczepanski, J., Amigo, J.M., Wajnryb, E., Sanchez-Vives, M.V.: Application of Lempel-Ziv complexity to the analysis of neural discharges. Network: Computation in Neural Systems 14, 335–350 (2003)
14. Lempel, A., Ziv, J.: On the complexity of an individual sequence. IEEE Transactions on Information Theory IT-22, 75–88 (1976)

The Effects of Spatial Attention on Face Processing: An ERPs Study

Liang Zhang* and Kan Zhang

Key Laboratory of Behavioral Science, Institute of Psychology,
Chinese Academy of Sciences, Beijing 100101, China
zhangl@psych.ac.cn

Abstract. Faces are one of the most biologically and socially significant objects in the human environment, and therefore believed to be processed automatically. To investigate whether the processing of faces is modulated by attention, event-related potentials (ERPs) were recorded in response to the dynamic facial stimuli. Spatial attention were manipulated by directing participants which location to be attended. The results showed that face-sensitive component N170 was not influenced by spatial attention, suggesting that the processing of faces was not modulated by the attention during the early stage. But the late-latency components were influenced by spatial attention. It indicated that the automatic processing of faces is more like to be partial rather than complete. These findings on dynamically real face processing by ERPs are expected to be used in the development of human-computer interactions.

Keywords: Face processing, Spatial attention, ERPs.

1 Introduction

Faces are one of the most salient objects for human survival and successful social interactions, as they convey essential information regarding identity, emotional expressions and intention. The PET and fMRI studies that focused on the faces processing over the past twenty years have been found the face-specific regions in the brain. Regions in the ventral visual pathway, particular the middle fusiform gyrus ("fusiform face area," FFA) have been shown to respond more strongly to faces than other objects [1-5].

These neuroimaging studies concerned the localization of faces processing. Event-related potentials (ERPs) provide a useful tool to investigate the time course of the processing because of the excellent temporal resolution [6]. The ERPs method has been widely used and found distinctive patterns of neural activity associated with faces processing, such as the face-sensitive component N170 [7-10]. The N170 is a negative component recorded from the posterior lateral electrode sites. It peaks at about 160–170 ms following stimulus onset and is recorded between 130 and 200 ms. This component is larger when elicited by human faces than by other object

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 541–550, 2013.

categories [11]. Also, a response component that occurs around 170 ms (M170) after stimulus onset has been described by previous studies with magnetoencephalographic (MEG) [12-14].

Based on the unique behavioral and physiological responses elicited by faces, many researchers concluded that faces are processed qualitatively distinctly from the other types of objects and by an anatomically well-localized modular system that is highly specialized for analyzing faces [15]. In this context, it is interesting to note that a number of studies reported that the processing of faces, which is unique and unlike other objects, appeared to be impervious to the influence of attention [16-18]. However, this conclusion was challenged by some studies [10, 15, 19]. These studies showed the opposite results that the faces processing was partially or completely modulated by attention. For example, Furey, et al. [14] investigated attention modulation on processing of faces with fMRI and MEG. The fMRI results showed that the response in the fusiform gyrus was strongly suppressed when attention was directed away from faces. However, the MEG results showed that attention had no effect on the M170, but late (>190 ms) category-related MEG responses elicited by faces were strongly modulated by attention. These conflict results may reflect that attention can modulate perceptual processing of faces at multiple stages. So far whether the processing of faces is automatic, or rather modulated by attention at early stage, is still debated.

Another issue must be considered is that static faces were used as stimuli in most of the faces studies (with only a few exceptions [20-22]), although dynamic facial expressions are encountered much more often than static facial "pictures" in everyday life. Despite the wide interest in the neural mechanisms of attentional modulation on faces processing, few study has been collected with dynamic facial stimuli. Neuroimaging studies have revealed that the brain regions known to be implicated in the processing of expressions, such as the posterior superior temporal sulcus (pSTS), the amygdala and the insula, respond more to dynamic than to static faces [23-25]. More importantly, authors reported cases of neurologically affected individuals that were incapable of recognizing static facial expressions but could recognize dynamic expressions [26].

The present study aimed to investigate whether the spatial attention influence the processing of faces, and more importantly, which stages of processing are modulated. It was manipulated by directing participants which location (either left or right side) is task relevant. Dynamic facial stimuli were employed in the study, because (1) neuropsychological and behavioral studies with normal people and patients revealed that the static and dynamic faces are processed differently [22-24]; (2) dynamic facial expressions come nearer everyday life than static faces.

Since the time course of the processing and the processing of ignored stimuli are of particular interest, event-related potentials (ERPs) were recorded. We hypothesized that (1) multiple stages of faces processing are modulated by spatial attention; (2) the influences of attention on different stages (early low-level stage, and late high-level stage) are different.

2 Methods

Participants. Fifteen healthy adults took part in the experiment. One participant had to be excluded from the data analysis because of excessive ocular and muscle artifacts in the electroencephalographic (EEG) recordings. The remaining 14 participants (9 female, 5 male, age: 23-29 years, mean: 26 years) were all right-handed, had normal or corrected-to-normal vision, normal hearing and had not reported any neurological disorders. Participants received monetary compensation for participation.

Stimuli. Human facial videos and voices were employed. Four professional actors (two male and two female) uttered nonsense pseudo words. Each actor's utterance was videotaped in a sound-attenuated recording studio and then converted to digital format. Actors were filmed from the frontal view and only their faces were shown with covered hair. Videos were presented gray scaled with a black background. Lighting, background, height, distance were controlled.

Videos were digitally sampled at 33 frames per second with 24-bit resolution at 640*480 pixel size. The auditory stimuli were the voices of the pseudo words. The audio tracks of the voices were equated for root mean square at 0.025 and digitally sampled with 16-bit. Mean duration of the stimuli was 700 ms (SE = 26 ms, ranged from 666 to 733 ms).

Procedure. Participants were seated in a sound attenuated and dimly lit room. They were in front of a computer screen. The viewing distance was maintained constant at 70 cm by using a chinrest. Visual and auditory stimuli were presented with an equal probability and in a random sequence on the left and right side. The visual angle of the visual stimuli was 3.4°× 4.9° (eccentricity=4.2°, measured as the distance between the centre of each face and the central fixation cross).

Participants' task was to attend to one modality (face or voice) of one side (left or right) and respond to the deviant stimuli of that modality at that side only (i.e., deviant stimuli of the attended modality and the attended location). The deviant stimuli differed from the standards in that they were inserted with a 200-ms interruption at the end of the videos/audios.

Only the data from face-attended trials were analysis in the present article. The data of crossmodal spatial attention was not analysis here.

Two "attend left faces" and two "attend right faces" sessions with 88 trials (72 standards and 16 deviants) each were presented. Additionally, one or two practice sessions were performed to familiarize participants with the task.

Each trial started with a central fixation cross (see Figure 1). After 400ms, a stimulus was presented on the left or right side of the fixation. Participants were asked to respond as accurately and as quickly as possible. All stimuli were presented in a random order with a stimulus onset asynchrony (SOA) of 1800-2100ms.

EEG recording. The EEG was recorded from 73 Ag/Ag-Cl-electrodes, mounted with equal distance into an elastic cap (Easy Cap; FMS, Herrsching-Breitbrunn, Germany). Electrodes were referenced to the right earlobe and re-referenced off-line to a linked earlobe reference (an averaged right/left ear lobe reference was calculated offline

using an additional left ear lobe recording). Vertical eye movements were measured with an electrode placed under the right eye, recorded against the right earlobe. Horizontal eye movements were monitored with a bipolar recording of two electrodes attached to the outer canthi of the eyes.

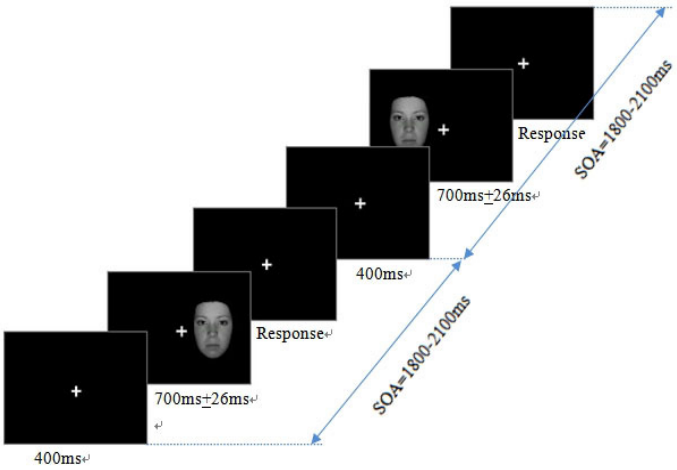


Fig. 1. The procedure used in the experiment, showing the sequence of events with two consecutive trails.

Electrode impedance was kept below 5 kΩ for scalp electrodes and below 10 kΩ for eye electrodes. The bandpass of the amplifiers (Synamps amplifiers; Neuroscan, Sipplingen, Germany) was set to 0.1–100 Hz and the digitization rate was 500 Hz.

Data analysis. EEG data processing was conducted with Vision Analyzer 1.0 (Brain Products GmbH) and included segmentation of the continuous signal into bins of 200 ms pre-, and 1000 ms post-stimulus. Only segments of standard stimuli were processed. Segments including a response and segments directly following a response were excluded from analyses. Only data to the visual stimuli were presented here, as the current question was to investigate the processing of faces.

The ERPs to the visual standard stimuli were averaged separately for each participant and attention condition, and referred to a 200 ms pre-stimulus baseline. Segments whose EOG activity exceeded 80 μV as well as segments with maximal amplitude differences exceeding 160 μV at any channel were rejected. Data from a participant were not used if more than 30% of the trials were discarded in this manner.

Electrodes were remapped to ipsilateral (i) and contralateral (c) recording electrodes with respect to the side of stimulation, and ERPs to stimuli from the left and right sides were pooled when they were attended and unattended, respectively.

For the statistical analyses, mean amplitudes were calculated for the selected time windows: P1 (100–140 ms), N170 (160–200 ms), P300 (340-500 ms) and a late

component (LC: 600-900 ms). Six electrodes were used in the statistical analysis where the P1, N170 and the late components were maximal in the right and left posterior regions (left: TPi, Pi, POi; right: TPc, Pc, POc). Time epochs and electrode sites were selected on the basis of the earlier studies (see [6, 9]) and on a visual inspection of the group grand average ERPs.

Three-way ANOVAs with factors Spatial attention (attended location vs. unattended location), Hemisphere (ipsi- vs. contralateral to the stimulation) and Electrode were run for spatial attention effect. Whenever the interaction with an Attention factor was significant, post-hoc tests for single electrode were calculated with paired t-tests (one-tailed).

Statistics were computed with SPSS, subroutine GLM for repeated measurements. Huynh/Feldt-corrected *P* values are reported where appropriate.

3 Results

3.1 Behavioral Results

The mean reaction times to the target visual stimuli measured from stimuli onset were 1399.6 ms (SE=44.4 ms). A one-way ANOVA (left vs. right) revealed that the accuracy was not influenced by Side of stimulation ($P>0.10$). Participants detected 92.6% (SE=1.7%) of the target faces. False alarms and misses rates were both below 1%. Side of stimulation (left vs. right) had no effect on accuracy, misses and false alarms (all $P>0.10$).

3.2 ERP Results

Visual stimuli elicited an early positive potential (P1) over the parietal and occipital areas, followed by a negativity peaked around 170 ms post-stimulus (N170) with a maximum over lateral parietal-occipital areas (See Figure 2). ANOVAs with three factors (Spatial attention × Hemisphere × Electrode) were run on the mean amplitudes of each interval. The ANOVA results were presented in Table 1.

P1 In this early time window, the ERPs to the location attended stimuli (thick lines in Figure 2) were more positive than ERPs to the location unattended stimuli (thin lines in Figure 2). The P1 effect was found [main effect of spatial attention: $F(1,13)=19.44$, $P=0.001$]. Post hoc tests showed that the spatial attention effect was most pronounced at bilateral parietal-occipital sites (Pi, POi, Pc and POc, all $P<0.01$). The effect was also found at temporal-parietal sites (TPi and TPc, $P<0.05$).

N170 No effect or interaction was found for the mean amplitude of N170, suggesting that this component was not influenced by spatial attention [spatial attention effect: $F(1,13)=2.64$, $P=0.128$].

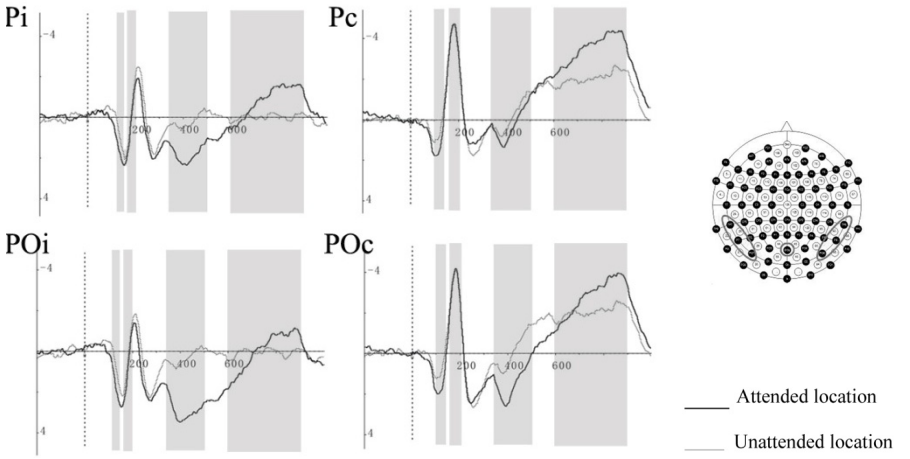


Fig. 2. Grand-averaged ERPs elicited by visual standard stimuli at parietal-occipital electrode sites (ipsilateral: Pi, POi; contralateral: Pc, POc). Time windows used in the statistical analyses are marked gray. Negativity is up.

Table 1. Results of ANOVAs for spatial attention effects

	P1	N170	P300	LC
Spatial attention (SA)	(1,13) = 19.44;.001	(1,13) = 2.64;.128	(1,13) = 3.44;.087	(1,13) = 8.02;.014
SA×Hemisphere	(1,13) = 0.01;.929	(1,13) = 0.19;.673	(1,13) = 10.64;.006	(1,13) = 5.65;.033
SA× Electrode	(2,26) = 6.60;.020	(2,26) = 0.03;.935	(2,26) = 10.96;.004	(2,26) = 4.71;.040
SA× Hemisphere×Electrode	(2,26) = 0.55;.508	(2,26) = 0.31;.622	(2,26) = 0.38;.573	(2,26) = 0.81;.807
Post hoc t-tests (Electrode)	TPi, Pi, POi, TPc, Pc, POc		Pi, POi, POc	Pi, POi, Pc, POc
Post hoc t-tests (*P*)	All *P*<0.05	All *P*>0.1	.007,.007,0.042	.001,.036,.001,007
F(*df*); *P*				

P300 In the time window 340-500 ms, ERPs to the location attended stimuli were more positive than ERPs to the location unattended stimuli. The main effects of Spatial attention were not significant but the interactions with Spatial attention were significant. The effects were not pronounced equally at all sites [spatial attention × Electrode: F(2,26) =10.96, P =0.004] and not equally over the two hemispheres [spatial attention × Hemisphere: F(1,13)= 10.64, P=0.006]. Post hoc tests showed a positive enhancement at Pi, POi (P<0.01), and POc (P=0.042). The tests also showed that the attention effects were more pronounced over the ipsilateral sites than the contralateral sites.

LC For this late time window, widely spatial attention effects were seen. The main effect of spatial attention was significant [F(1,13)= 8.02, P=0.014]. This effect was not pronounced equally at all sites [spatial attention × Electrode: F(2,26) =4.71, P

=0.040] and not equally over the two hemispheres [spatial attention × Hemisphere: F(1,13)= 5.65, P=0.033]. Post hoc tests showed a negative enhancement at bilateral parietal and occipital sites (Pi, POi, Pc, POc, all P<0.05).

4 Discussion

The aim of the present study was to investigate whether the processing of faces is affected by spatial attention. ERPs were recorded in response to the dynamic facial stimuli presented in spatial attended/unattended condition.

The results can be summarized in four points. First, the spatial attention effect started at about 100 ms after stimulus onset, as indexed by a reduction of the P1 component to the stimuli of the unattended location. Second, the N170 was impervious to the influence of spatial attention. Finally, spatial attention modulated the ERPs response during the late stages (as indexed by P300 and LC). The implications of these results on the way the brain processes faces were discussed.

Is face processing automatically?

In the present study, participants had to detect infrequent target stimuli at the attended location only, while ignoring stimuli at the unattended location. However, both of the attended and unattended faces elicited the face-sensitive component N170. Moreover, this component appeared to be unaffected by attention, which is assumed to reflect the encoding of faces was not modulated by attention. In contrast, P1 amplitude is influenced by attention modulation. By examining P1 amplitude, we confirmed that the absence of N170 modulation reflects the specialty of faces processing. These results are consistent with previous works in which early face responses were not modulated by attention [16-17, 27].

One view that has been invoked to explain the absence of early attentional effects for faces is that faces enjoy a privileged status and are fully processed automatically [16, 28]. However, P300 effect and the late negative effect were observed in our study, indicating the faces are not fully processed automatically.

A similar result pattern has been found by a MEG study. Lueschow, et al. [17] found that the first face-distinctive MEG response was observed at 160-170 ms (M170). Nevertheless, attention did not start to modulate face processing before 190 ms. Such results suggest although face processing may be impervious to attentional influences during early stages of sensory perceptual process, attention may influence later face processing. From our ERPs results, the spatial attention effects at late components indicate that obligatory processing may occur for “ignored” faces, but it is more like to be partial rather than complete.

How early are spatial modulation and faces encoding? The difference between dynamic and static faces.

The results showed that attended faces versus unattended faces produced an early modulation of posterior ERPs components from 100 ms after stimulus onset (P1). It is well established that early sensory processing, as indexed by the P1 component, is

modulated by spatial attention [29]. The studies with faces stimuli also reported that spatial attention modulated visual processes as early as about 80 ms [30]. Whereares, the spatial attention effect eliminated after 140 ms and did not appeare at the face-sensitive component N170. One possible explanation is that the face encoding started at about 160 ms post stimulus (N170). This finding is similar with the ERPs results of Jacques & Rossion [30]. They empolyed two concurrent faces and modulated spatial attention. It was reported that the spatial attention modulation was earlier than the representations of the concurrent faces. The neural representations of the concurrent faces was in occipital-temporal cortex as early as 130 ms. These results may suggest that spatial attention modulates visual processes as early as about 100 ms after stimulus onset and the face encoding take place around 140 ms post stimulus.

However, there was ERP results suggested that faces were categorized around 100 ms [31]. The difference between these results and the current results may due to the nature of dynamic faces comparing to static faces. A PET study has been proved that encoding of facial expressions by static or dynamic displays is associated with different distributed network of brain regions. Differential activation of visual area V5, superior temporal sulcus, periamygdaloid cortex involved in the processing of dynamic and static expressions. But considering the various types of stimuli, paradigms, and measurements, we cannot draw the conclusion that the encoding of static and dynamic faces takes place differently. Further research should compare the time course of static versus dynamic faces processing under the same condition.

Perceptual load modulation of early face processing.

Perceptual load has been shown to influence early sensory processing [32]. Two recent ERPs researches argued that the early stages of face processing indexed by the N170 strongly suppressed by perceptual load manipulation [30, 33]. According to the perceptual load theory of selective attention (reviewed by Lavie[34]), the irrelevant stimulus should be processed when spare capacity can “spill over” to irrelevant items (i.e., low attentional load), but not when all available capacity are exhaust (i.e., high attentional load). Another recent research proposed that the attentional selection effects were influenced by face discriminability during attention to faces. They found that early selection (N170 modulation) was present for low and medium, but not for high discriminability faces, whereas selection at a later stage was comparable for all levels of discriminability [10]. The results can also be interpreted as the perceptual load manipulation. Comparing to the low and medium discriminability faces, the high discriminability faces were at low attentional load so that N170 was suppressed. Although we did not manipulate the levels of perceptual load, comparing to the previous results, a plausible account for the absence of N170 modulation on faces in our study is that perceptual load remained at a low level in our experiment. Under such condition, the capacity was not exhaust so that the irrelevant faces were processed by the spare capacity. It has to be noted that the faces in Neumann and Sreenivasan’s studies were both at the center of fixation. The selective attention were manipulated by task demands. The mechnism of the selective attention could be different from the spatial attention of the present study in which the attention was manipulated by transfer the focus of attention.

Besides, the previous studies and our study are consistent in the ERPs results at late stages. These results support that the late stage of faces processing are influenced by attention.

In summary, the results bring further support to the view that faces are processed rapidly and independently of attentional modulation during early perceptual stage. But attention can influence the later dynamic face processing stage.

Acknowledgements. This work was supported by grants from the National Natural Science Foundation of China (No.31100734 and No.91124003), and the National Basic Research Program of China (973 Program, No 2011CB711000). The authors would especially like to thank Prof. Brigitte Roeder (University of Hamburg) and her lab.

References

1. Puce, A., et al.: Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. Journal of Neuroscience 16(16), 5205–5215 (1996)
2. Kanwisher, N., McDermott, J., Chun, M.M.: The fusiform face area: A module in human extrastriate cortex specialized for face perception. Journal of Neuroscience 17(11), 4302–4311 (1997)
3. Betts, L.R., Wilson, H.R.: Heterogeneous Structure in Face-selective Human Occipito-temporal Cortex. J. Cogn. Neurosci. (2009)
4. Cohen Kadosh, K., et al.: Task-dependent Activation of Face-sensitive Cortex: An fMRI Adaptation Study. J. Cogn. Neurosci. (2009)
5. Liu, J., Harris, A., Kanwisher, N.: Perception of Face Parts and Face Configurations: An fMRI Study. J. Cogn. Neurosci. (2009)
6. Eimer, M., Driver, J.: Crossmodal links in endogenous and exogenous spatial attention: evidence from event-related brain potential studies. Neurosci. Biobehav. Rev. 25(6), 497–511 (2001)
7. Botzel, K., Schulze, S., Stodieck, S.R.G.: Scalp Topography and Analysis of Intracranial Sources of Face-Evoked Potentials. Experimental Brain Research 104(1), 135–143 (1995)
8. Bentin, S., et al.: Electrophysiological studies of face perception in humans. Journal of Cognitive Neuroscience 8(6), 551–565 (1996)
9. Eimer, M.: The face-specific N170 component reflects late stages in the structural encoding of faces. Neuroreport 11(10), 2319–2324 (2000)
10. Sreenivasan, K.K., et al.: Attention to faces modulates early face processing during low but not high face discriminability. Atten. Percept. Psychophys. 71(4), 837–846 (2009)
11. Rossion, B., Jacques, C.: Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? Ten lessons on the N170. Neuroimage 39(4), 1959–1979 (2008)
12. Halgren, E., et al.: Cognitive response profile of the human fusiform face area as determined by MEG. Cerebral Cortex 10(1), 69–81 (2000)
13. Liu, J., Harris, A., Kanwisher, N.: Stages of processing in face perception: an MEG study. Nature Neuroscience 5(9), 910–916 (2002)
14. Furey, M.L., et al.: Dissociation of face-selective cortical responses by attention. Proc. Natl. Acad. Sci. U S A 103(4), 1065–1070 (2006)

15. Crist, R.E., et al.: Face processing is gated by visual spatial attention. Frontiers in Human Neuroscience 1 (2008)
16. Cauquil, A.S., Edmonds, G.E., Taylor, M.J.: Is the face-sensitive N170 the only ERP not affected by selective attention? Neuroreport 11(10), 2167–2171 (2000)
17. Lueschow, A., et al.: Looking for faces: Attention modulates early occipitotemporal object processing. Psychophysiology 41(3), 350–360 (2004)
18. Finkbeiner, M., Palermo, R.: The role of spatial attention in nonconscious processing: a comparison of face and nonface stimuli. Psychol. Sci. 20(1), 42–51 (2009)
19. Brassen, S., et al.: The influence of directed covert attention on emotional face processing. Neuroimage (2009)
20. de Gelder, B., et al.: The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. Neuroscience Letters 260(2), 133–136 (1999)
21. Kreifelts, B., et al.: Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. Neuroimage 37(4), 1445–1456 (2007)
22. Collignon, O., et al.: Audio-visual integration of emotion expression. Brain Research 1242, 126–135 (2008)
23. LaBar, K.S., et al.: Dynamic perception of facial affect and identity in the human brain. Cerebral Cortex 13(10), 1023–1033 (2003)
24. Kilts, C.D., et al.: Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions. Neuroimage 18(1), 156–168 (2003)
25. Haxby, J.V., Hoffman, E.A., Gobbini, M.I.: The distributed human neural system for face perception. Trends in Cognitive Sciences 4(6), 223–233 (2000)
26. Adolphs, R., Tranel, D., Damasio, A.R.: Dissociable neural systems for recognizing emotions. Brain and Cognition 52(1), 61–69 (2003)
27. Carmel, D., Bentin, S.: Domain specificity versus expertise: factors influencing distinct processing of faces. Cognition 83(1), 1–29 (2002)
28. Lavie, N., Ro, T., Russell, C.: The role of perceptual load in processing distractor faces. Psychol. Sci. 14(5), 510–515 (2003)
29. Hillyard, S.A., Anllo-Vento, L.: Event-related brain potentials in the study of visual selective attention. Proc. Natl. Acad. Sci. U S A 95(3), 781–787 (1998)
30. Jacques, C., Rossion, B.: Electrophysiological evidence for temporal dissociation between spatial attention and sensory competition during human face processing. Cerebral Cortex 17(5), 1055–1065 (2007)
31. Pegna, A.J., et al.: Visual recognition of faces, objects, and words using degraded stimuli: where and when it occurs. Hum. Brain Mapp. 22(4), 300–311 (2004)
32. Eimer, M.: Crossmodal links in spatial attention between vision, audition, and touch: evidence from event-related brain potentials. Neuropsychologia 39(12), 1292–1303 (2001)
33. Mohamed, T.N., Neumann, M.F., Schweinberger, S.R.: Perceptual load manipulation reveals sensitivity of the face-selective N170 to attention. Neuroreport 20(8), 782–787 (2009)
34. Lavie, N.: Distracted and confused?: selective attention under load. Trends in Cognitive Sciences 9(2), 75–82 (2005)

Part V

Cognitive Load, Stress and Fatigue

The Information Exoskeleton: Augmenting Human Interaction with Information Systems

James P. Allen, Susan Harkness Regli, Kathleen M. Stibler, Patrick Craven, Peter Gerken, and Patrice D. Tremoulet

Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ
[james.p.allen,susan.regli,kathleen.m.stibler,patrick.craven,
peter.gerken,polly.d.tremoulet]@lmco.com

Abstract. In the military intelligence cycle the warfighter acts as both a receiver and a producer of information. As a receiver the warfighter must be able to readily assimilate disparate mission-relevant information. As a producer the warfighter must be cognizant of both the current information requirements and the ability to meet them. Both of these tasks are exacerbated by the heat of battle and, in the case of the receiver, the ever-increasing amount of available information. To address these challenges Lockheed Martin Advanced Technology Laboratories (LM ATL) is creating a suite of capabilities to augment warfighter interaction with intelligence services. Much like a powered exoskeleton augments human interaction with the physical environment, our Information Exoskeleton augments the warfighter's interaction with intelligence, providing greater situational awareness with minimal operational overhead. This paper describes our vision for the Information Exoskeleton, the capabilities required to realize it, and related research efforts.

Keywords: Information Exoskeleton, Information Needs Assessment, Context Awareness, Information Alignment, Cognitive Alignment.

1 Introduction

The past two decades have witnessed a dramatic rise in military intelligence collection and dissemination. Advances in electronics, communications, and automated technologies for performing data integration, analysis, and dissemination have made it possible to rapidly push increasing amounts of intelligence to warfighters. The recent proliferation of mobile devices means that dismounted warfighters are increasingly able to a) to receive intelligence in the field, and b) collect and disseminate tactical information essential to the generation of intelligence while supporting ongoing missions.

Actionable intelligence is key to success in tactical operations such as reconnaissance patrols, cordon and search or combat patrols, but its utility is undermined if it is not presented in a fashion that allows it to be easily understood and applied for greater situational awareness. Interacting with a myriad of information from different sources can impose significant cognitive and physical burdens (Claburn, 2009; Shanker &

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 553–561, 2013.

Richtel, 2011). Dismounted warfighters are forced to maintain the shifting operating picture, mostly in their heads, while taking into account data from multiple systems and devices such as Blue Force data, enemy position reports, audio communications, video and RF signal detection, requiring extensive context switching (Hsu, 2011). If intelligence is delivered to these warfighters without regard to timing, relevance, or modality it will likely be underutilized, or may overwhelm or distract them when lives are at stake.

Research on attention and multitasking suggests that in demanding situations requiring sustained attention, especially life-threating ones, individuals have difficulty successfully multitasking. This has been demonstrated in classroom learning conditions where students who were allowed to use laptops to browse and use social media during a lecture suffered decrements on tests of memory compared to peers who did not split their attention (Hembrooke & Gay, 2003), as well as a driver's ability to quickly respond to driving-related stimuli is hindered by either handheld or hands-free use of a cell phone (Horrey & Wickens, 2006) However, other research has shown that certain military functions like sentry duty allow a warfighter to successfully manage both the visual scan task as well as responding to auditory signals (McBride, Merullo, Johnson, Banderet, & Robinson, 2007). In fact, the researchers observed that when the work rate was increased, overall performance improved. These results suggest that attention is an important resource that cannot be overly taxed lest it result in delayed or missed reactions, or under-utilized lest it result in an individual tuning out from the task at hand. Technology to support the warfighter must take these extremes into account and carefully estimate the level of attention to help keep the warfighter in an optimal state to respond effectively to incoming intelligence while opportunistically collecting data relevant to known intelligence requirements.

What is needed to better equip dismounted warfighters for current and future operations is a system built upon a solid framework that supports the constantly changing needs of the warfighters and their shifting context. For the past decade, investigators at Lockheed Martin Advanced Technology Laboratories have been conducting research toward our vision of an Information Exoskeleton (IE) for the warfighter. Much like a powered exoskeleton augments human interaction with the physical environment, our Information Exoskeleton augments the warfighter's interaction with intelligence, enhancing the warfighter's ability to benefit as a consumer of mission-relevant information and also act as an intelligence producer with minimal operational overhead. The IE ensures that the intelligence cycle provides the greatest situational awareness (SA) with the least amount of operational disruption. This paper describes our operational vision of the IE, explores the challenges and required capabilities to enable it, and presents our current and planned research efforts.

2 Information Interaction with the Warfighter

2.1 Concept of Operations

To help convey the utility of the IE, we present a Concept of Operations where the IE assists a ground warfighter during his patrol mission.

A dismounted ground warfighter is preparing to go out on patrol in a dynamic urban environment. Prior to a patrol, the warfighter typically receives an intelligence briefing to specify what threats or other activities have occurred recently in the patrol area of operation. With the IE, the warfighter will also be outfitted with body-worn physiological sensors (e.g., monitors for heart rate, blood pressure, galvanic skin response) and tactical sensors (e.g., accelerometer, gunshot detection sensor, blast detection sensor, microphone, camera, gyroscope). The data from these sensors are wirelessly collected by a small handheld or wearable device that provides applications the warfighter can use to file digital intelligence reports, using either texting or spoken language understanding technology, and to receive updates to his understanding of the tactical situation while he is on patrol. The IE system will collect data from the warfighter to understand the warfighter's context of operation, including position, health status, engagement in combat activity, and in turn will use this understanding of individual context to help decide what intelligence to provide the warfighter, as well as the best way to present the information.

Fig. 1. IE enables the warfighter to efficiently process information

The warfighter provides the IE with a plan for the likely patrol, and the IE checks to see if there is applicable intelligence to the planned mission such as known threats, maps, weather information, and terrain. Once out on patrol, a nearby explosion occurs. As the squad responds with an immediate action drill the IE detects a pattern of inputs from the microphone, accelerometer and gyroscope that matches the signature of an improvised exploding device (IED). The IE initiates actions to aid in near-term tasks. Sound is being recorded on all devices in the squad and the IE requests that other sensors (e.g., GPS, accelerometer) begin logging data to capture movement and changes in posture. The IE creates an observation report template and enters current location and time information so the warfighter can complete and transmit the pre-populated report with information about the IED.

Continuing along the route, the warfighter receives an audio alert that the IE has intelligence that shows a black car blocking the planned route. Suspecting an ambush, the warfighter diverts to an alternate road. The IE prepares information that is relevant to the new route, but recognizes from the speed the warfighter is moving, and the increased stress indicated by physiological sensors, that the warfighter likely cannot attend to the relatively low priority new information. IE begins summarizing and filtering the information based on priority, and stores it for future delivery to the warfighter when cognitive load is lower and assimilating the information is possible. At this point the IE detects that the warfighter is in a good location to collect information to help satisfy a commander's information requirement, and generates an alert that will be sent along with the new route data when the warfighter is ready to receive it.

3 Capability Requirements

The dissemination of intelligence to a warfighter can greatly increase SA of the battlespace. However, pushing information without regard to timing and usability can negatively impact warfighters. Three major challenges that must be addressed to effectively disseminate intelligence to dismounted warfighters are 1) determining the relevance of information to the warfighter given the dynamics of the battlespace, 2) ensuring the usability of available intelligence and 3) effectively presenting information to that warfighter. Satisfying these will enable the warfighter to achieve the highest level of SA with the least amount of operational disruption.

Warfighters must be cognizant of the commander's intelligence requirements as they go into battle because at any given time they may observe, or discover, new information that can strengthen the commander's SA or satisfy the existing requirements. There are challenges for the tactical warfighter to overcome in order to collect the right information. The tactical warfighter needs knowledge of the requirements along with help recognizing when to collect information and the facility to capture information essential to the generation of intelligence. The goal for collection of useful tactical information is to maximize SA while minimizing operational overhead.

We believe that the challenges to interaction for both information dissemination and collection can be addressed by three core IE capabilities:

1. Assessing the warfighter's operational context
2. Assembling information based on context
3. Adapting the user interface to the information and user operational context

While engaged in an activity, individuals are in a particular cognitive state and exhibit predictable physical conditions. These cognitive and physical indicators are part of the tactical warfighter's operational context. The context consists of elements such as geographic position, health status and engagement in a combat activity. It can be thought of as a plan or task being executed by a tactical warfighter along with physical and cognitive states. Physical context attributes include physiological response and body position, while cognitive state describes individual awareness, cognitive load, and current interests. Body-worn physiological and tactical sensors assess heart

rate, blood pressure, pulse ox, stature, detect and geo-locate signal activity (Regli, Tremoulet & Stibler, 2013). The operational context may also include descriptions and status of his current mission, his role in the mission, and environmental details.

Currently, the volume of information being processed by Companies is large enough to justify a dedicated Company Intelligence Support Team (CoIST) which assists the commander in intelligence analysis and fusion, else the intelligence becomes stale and dangerously obsolete to patrols (Morgan, 2008). While there is a wealth of data being collected and fashioned into intelligence, very little of it is ever used by the tactical warfighter. The reason is two-fold. The tactical warfighter does not have the bandwidth to scrutinize data and correlate it with other information and is not able to constantly monitor a screen while patrolling with a weapon in hand. While just providing tactical warfighters with all possible intelligence seems to be the solution, it would create larger problems by disrupting their primary task and information overload. Tactical warfighters cannot spend time sorting through data. They need correlated sets of information relevant to their current task and environment verses streams of information. They can draw some conclusions in the field, but really need someone to help them "connect the dots" to have a greater situational understanding of the battlefield in which they are operating. Filtering data and correlating information from various sources will ensure that the warfighter receives more manageable amounts of highly relevant data. For example, the warfighter might be interested in historical IED blasts along his mission route, but only those that have occurred within a pre-defined timeframe. Another way to reduce the data would be to determine a pattern in the blasts. Maybe they occur at a particular time in the day. Maybe they are triggered by another event such as the passing of a convoy through an intersection along the route. The capability to provide tactical warfighters with controlled amounts of relevant data will enhance their situational awareness while still allowing them to successfully perform their primary mission.

Timely intelligence is most beneficial to warfighters when it is delivered via a method that enables them to rapidly assimilate the information, thus minimizing disruption from primary tasks. Relevancy is vital since context switching is extremely difficult and potentially dangerous in their operational environment. Presentation of the information is equally important. The most appropriate set of modalities (visual display, auditory or tactile alert) for presenting new intelligence depends upon warfighter context, including the immediate environment (noisy? potentially threatening? light sensitive?) and what tasks are being performed (patrolling an area? looking for a specific vehicle?). For successful information transfer, tactical warfighters need a system that has the capability to adapt its timing and communication modality to the user's tasks and environmental constraints.

Establishing a contextual understanding of the user allows for collection and dissemination of information relevant to that context. A warfighter's operational context leverages the most current information, correlates it with the known information and incorporates it into an existing perspective. Since the system has been tailored to present only information relevant to a warfighter's mission, including the current location and route, we expect a reduction in review time. Before sending data, the IE verifies that the warfighter is in the right context to be able to process the data. If the

physiological and cognitive sensors suggest an overload and the intelligence is not time-sensitive and does not pose an immediate threat to the warfighter, it may wait to notify him. It will also refrain from detailing existing information requirements and identifying opportunities for collection of information that may satisfy those requirements. Pertinent information is presented to warfighters using formats and modalities that make it easy for them to understand how the information is relevant to their missions.

4 Component Structure and Applicable Research

Figure 2 presents a diagram that shows how the main IE components work together to provide information to the user. This section will focus on the function of each component and the areas of investigation we have conducted to lay a foundation for our IE vision. In addition to a reference structure, we have created prototype implementations on Android platforms to enable active engagement with subject matter experts to help evaluate and enhance the capabilities.

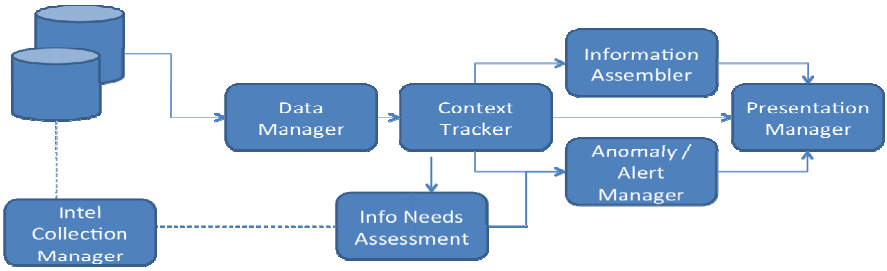


Fig. 2. Information Exoskeleton Component Diagram

The ***Data Manager*** stores all of the information collected by the system in a form that is easily accessible from the rest of the system. Data can come from sources varying from biometric sensors, GPS, accelerometers, to reports, blue force tracking and RF detections.

To collect data about the warfighter's context without having to interrupt the warfighter's activities, we developed a generic data-modeling component to provide automated collection, storage, and dissemination of sensor data generated by body-worn physiological sensors and/or other tactical sensors. When data is requested from another warfighter, messages are constructed with the sensor outputs and sent over the network to that warfighter's device. We have demonstrated collecting sensor data from simulated sensors and from real sensors that transmit their sensor data via Bluetooth wireless communication. This sensor technology has been applied in several domains, including medical triage and casualty reporting.

As information is collected, multiple trackers and components work together to organize the information being received. The ***Context Tracker*** determines what is relevant to the human based on a set of predetermined contexts: biometrical, world, human, and tactical contexts.

We have developed prototype systems that adapted the user interfaces based upon user task and operational contexts. As part of these efforts we developed the abilities 1) to assess individual's cognitive states through physiological data, and 2) to track and manage the tasks that require operator attention (Morizio, Thomas, and Tremoulet, 2005). We developed mitigation strategies to minimize disruption of the user's primary task (Regli, Tremoulet, Hastie, and Stibler, 2006). More recent research has involved tracking and adapting user interaction based on other aspects of warfighter context such as mission status, walking versus stationary, etc.

Our research in Plan Execution Monitoring (Allen, McCormick, 2005) enables the IE to be intelligently informed about the changing nature of the operational context. The plan monitor compares the values returned from environmental sensors with the values in the models to determine which activity is currently being executed, and the status of that activity. For military domains where explicit plans are used (e.g., tactical missions) the IE can leverage this approach to determine plan state allowing the context tracker to know the warfighter's current activity. Armed with this contextual information the **Information Assembler** is better equipped to provide mission-specific information for the warfighter.

The ***Anomaly/Alert Manager*** component monitors the data and produces alerts based on the rules of the context in which it is operating.

Our Human Alerting and Interruption Logistics - Surface Ship (HAIL-SS) system is based on anomaly monitoring and alert management research. HAIL provides alert management to maximize the benefit of timely critical alerts and minimize negative effects of human interruption. It is composed of services that alert human operators appropriately, and help the operators recover work-flow context afterwards. HAIL-SS enables operators to maintain higher levels of situational awareness despite a high volume of alerts that are generated from automation. (McFarlane, 2006)

Alerts generated by the ***Alert Manager*** are delivered to the ***Information Assembler*** for processing. The ***Information Assembler*** collects, organizes and correlates relevant data, producing useful information for the warfighter.

We have been investigating how to most effectively express information that is typically requested and used by intelligence analysts in a manner that is consistent with the tactical language and perspective of a warfighter on patrol. The first part of the challenge is enabling queries to be expressed in tactical language by presenting tactical vocabulary as a front end to queries that contain logic gleaned from intelligence experts (Samoylov et at., 2009). It also requires an understanding of tactical tasks to enable the presentation of different types of data from multiple data sources in a manner that is correlated and filtered to match the task goals that the warfighter is trying to accomplish. This area of information assembly research is ongoing.

The ***Presentation Manager*** component determines how to deliver the data via an appropriate set of modalities to the user's interface based on the user's current needs.

We developed an "environment director" component that selects presentation modalities based on the task's preferred modality, the application's modality capabilities, and user context. More recently, to enable geographic display of relevant information including blue force tracking and reports by location, we have developed a lightweight graphics library that can display geo-rectified objects (e.g. map tiles, icons, grid lines) on a map display. The library supports panning and zooming in and out of the map tiles that are stored at different zoom levels. The current GPS position

of each device is collected and shared with all the other devices and shown on the map. We have employed this lightweight mapping and visual location display technology on several efforts, including the observation reporting domain and medic triage and casualty reporting domain.

In the dynamic battlespace information requirements change along with the operational context. Given updated contexts from the context tracker, the ***Information Needs Assessment*** component determines the types of information that best support the current state of the mission. We have conducted extensive research into automatic and semi-automatic approaches to anticipate the information needs of the warfighter. Our fully automated approach is based on direct mapping of mission state and warfighter role to information requirements based on historical/statistical analysis of prior information requests. Our semi-automated approach is a recommender system that leverages information ontologies, historical analysis of prior information requests, and mission state. When the warfighter requests information the system recommends additional information that might be relevant.

Finally, the ***Intel Collection Manager*** supports the warfighter by providing simple, intuitive, multi-modal interfaces for gathering and disseminating tactical intelligence.

Our capability in the area of tactical information collection in support of intelligence generation enables the warfighter to speak the contents of standard tactical reports; the spoken utterances are parsed into structured digital reports that can be shared more easily, sent to a tactical operations center (TOC) when possible, and made available for use by other warfighters or by intelligence analysts for near- or long-term increase in overall situational awareness of the battlefield. A multimodal interface enables report entry by voice when hands are occupied and by text when there is a need to remain quiet. We have applied spoken language understanding technology to several domains including small unit logistics, squad-level observation reporting and casualty reporting.

5 Discussion

Our research is guided by a vision of an intelligent Information Exoskeleton that seamlessly allows the right information to be collected, processed, pre-positioned, requested, and delivered in a manner that amplifies human effectiveness. The IE hosts a suite of capabilities that understands a user's tasks, anticipates needs, assists in gathering knowledge and presents relevant information in a time, format and modality-appropriate way that minimizes disruption. As such, the IE functions as a contextual window between a tactical user and the world.

Our future efforts will focus, primarily, on enhancing this contextual window by expanding and aggregating our views into the warfighter's operational context. While we are currently able to monitor executing missions, some roles in the military aren't represented by such explicit, well-defined plans (e.g., intelligence analyst). Recent research in Task Context Management (Kersten & Murphy, 2006) shows that such tasks can be tracked with minimal, if any, human intervention. Additionally, we would like to develop techniques for aggregating the disparate contexts into a unified warfighter profile that can be leveraged by both the IE and other information systems to provide better intelligence to the warfighter.

References

1. Allen, J.P., McCormick, J.M.: Adaptive Plan Monitoring Systems for Military
2. Decision Support. Challenges to Decision Support in a Changing World, Papers from the, AAAI Spring Symposium, Technical Report SS-05-02, March 21-23, AAAI Press, pp. 1–2 (2005)
3. Claburn, T.: Military Grapples With Information Overload. Information Week (2009), http://www.informationweek.com/government/enterprise-architecture/military-grapples-with-information-overl/218401332
4. Hembrooke, H., Gay, G.: The laptop and the lecture: The effects of multitasking in learning environments. Journal of Computing in Higher Education 15(1), 46–64 (2003), doi:10.1007/bf02940852
5. Horrey, W.J., Wickens, C.D.: Examining the Impact of Cell Phone Conversations on Driving Using Meta-Analytic Techniques. Human Factors: The Journal of the Human Factors and Ergonomics Society 48(1), 196–205 (2006), doi:10.1518/001872006776412135
6. Hsu, J.: Military faces info overload from robot swarms. NBC News (2011), http://www.nbcnews.com/id/44430826/ns/technology_and_science-innovation/
7. Kersten, M., Murphy, G.C.: Using Task Context To Improve Programmer Productivity. In: Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering, pp. 1–11. ACM, New York (2006)
8. McBride, S.A., Merullo, D.J., Johnson, R.F., Banderet, L.E., Robinson, R.T.: Performance During a 3-Hour Simulated Sentry Duty Task Under Varied Work Rates and Secondary Task Demands. Mil. Psychol. 19(2), 103–117 (2007), doi:10.1080/08995600701323392
9. McFarlane, D.C.: Engaging Innate Human Cognitive Capabilities to Coordinate Human Interruption: The HAIL System. In: Forsythe, C., Bernard, M.L., Goldsmith, T.E. (eds.) Forsythe, Chris; Bernard, Michael L.; Goldsmith, Timothy E, pp. 2137–2152. Lawrence Erlbaum Associates Publishers, Mahwah, x, 2314 (2006)
10. Kersten, M., Murphy, G.C.: Using task context to improve programmer productivity. In: Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering, Portland, Oregon, USA, November 5-11 (2006)
11. Morgan, R.: Company Intelligence Support Teams. Armor, 23–50 (July-August 2008)
12. Regli, S., Tremoulet, P., Stibler, K.: C4ISR-Med Battlefield Medical Demonstrations and Experiments, January 2012- February 28, 2013. Lockheed Martin Advanced Technology Laboratories (2013), http://www.atl.external.lmco.com/library.php
13. Regli, S.H., Tremoulet, P., Hastie, H., Stibler, K.: Mitigation Strategy Design for Optimal Augmented Cognition Systems. In: Schmorrow, D.D., Stanney, K.M., Reeves, L.M. (eds.) Foundations of Augmented Cognition, Proceedings of the 2nd annual Augmented Cognition Conference, San Francisco, CA, October 15-20 (2006)
14. Samoylov, A., Franklin, C., Regli, S.H., Tremoulet, P., Stibler, K., Gerken, P.: Tactical Access to Complex Technology through Interactive Communication (TACTIC). In: Smith, M.J., Salvendy, G. (eds.) HCI International 2009, Part I. LNCS, vol. 5617, pp. 154–162. Springer, Heidelberg (2009)

QEEG Biomarkers: Assessment and Selection of Special Operators, and Improving Individual Performance

Donald R. DuRousseau

PEAK Neurotraining Solutions, Sterling VA, U.S.A.
don@peaknt.com

Abstract. Future military special operator selection and education programs will take advantage of state-of-the-art neuroimaging and normative statistical tools in the creation of a customized database of EEG patterns gathered from top performing specialists over their careers. Such a quantitative EEG Normative Database (qEND) will function as the benchmark for screening, assessment, selection and even training of targeted individuals required to work effectively as operators under extreme stresses and for extended periods. This assumption implies that an improved warfighter selection and training pedagogy will embrace the concept of a "model" brain activity pattern (BAP) that represents a warfighter at peak potential and in a highly focused and resilient state of mind. It also implies that this model BAP can be used to: 1) identify biomarkers of positive traits in candidates for specialized training programs, and 2) reduce stress and improve sleep and training performance of program selectees using guided EEG neurofeedback to maintain an optimal BAP. One such statistical qEND (NeuroGuide) is used clinically in the assessment and diagnosis of EEG imbalances specifically related to neurological and behavioral disorders, as well as for guiding individual brain pattern changes through the use of neurofeedback training (NT).

To evaluate qEEG for monitoring an individual's BAP changes and potentially improving mood and work performance, two military specialists with leadership experience underwent a program of pre- and post-EEG recordings and 20 neurofeedback training (NT) sessions. Here, the NeuroGuide database was used to determine how each participant's BAP differed from the age-matched group norms, and it was also used during the NT process to inform the software of the differences from the norms at each of the 4 training sites used to adjust the trainees EEG towards the direction of "normal".

Changes from the NT program were assessed pre- and post-intervention using seven neuropsychological assessments of mood, anxiety, sleep, work performance and life satisfaction. In addition, one subject had a series of blood draws taken over the course of the NT program to evaluate changes in his plasma Cortisol; a reliable biomarker of stress level. Both subjects reported reduced levels of anxiety, impulsivity and anger, and improved mood and life satisfaction after the 20-session NT intervention.

Keywords: Assessment and Selection, Biomarkers, Quantitative EEG, Neurofeedback, Normative Statistics, Training Technologies, Training Policy.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 562–571, 2013.

1 Introduction

Adult electroencephalography (EEG) patterns are individually stable with predictable age-related patterns of change in amplitude, coherence and phase measures of resting brain wave activity (1, 2). This predictability allows the use of comparative statistics in the evaluation of an individual's BAP and its comparison against an age-matched norm of over four-thousand independent EEG measures (3). What follows from this statistical measurement capability is the development of a normative database of EEG features gathered from a large number of "normal individuals". Such a database makes it possible to quantify the statistical differences in the brain waves of one person as compared to their age-matched group averages (4). Several EEG normative database products exist today for assessing individual brain imbalances (NX Link), prescribing psycho-active medications (Reference EEG), and as a neurofeedback training modality (NeuroGuide) primarily for individuals coping with anxiety, stress, insomnia, depression, addiction, obsessive compulsive disorders, cognitive difficulties and behavioral problems.

This paper attempts to describe an analytical approach known as the quantitative EEG Normative Database (qEND) and its use in assessing brain functions and guiding neurofeedback training (NT) protocols that may be used to reduce stress and enhance mental and physical resilience in the warfighter. There is a scientific basis for use of NT as a means to help combatants maintain peak levels of performance under stress, and a strategy exists for the rapid development, validation and deployment of enhanced neurotechnologies specifically targeting rapid expert level knowledge and skills acquisition. The use of EEG neurofeedback to normalize brain wave activity has consistently been shown to improve sleep and reduce anxiety and it is widely available all over the world today (5, 6, 7).

2 The Plastic Brain

One thing we can all agree upon is that chronic stress will change your brain, particularly in areas associated with memory, sleep and emotional regulation (8). Long-term stress changes hormone levels, which in turn modulate neurotransmitter production and uptake; driving lasting changes in the EEG (9). Over time, imbalanced EEG patterns reorganize in the brain's key system-level networks, ultimately establishing a new "yet stable" brain activity pattern (BAP) (10). With reinforcement, this imbalanced BAP can stabilize through a resonant process that perpetuates the thoughts and feelings associated with prolonged exposure to high stress; like anxiety, rumination, panic, depression, and contemplating suicide. Ultimately, this highly imbalanced BAP becomes the norm and repetitive patterns of negative or self-effacing behaviors develop and become linked to this now-stable imbalanced brain state (11). NT provides a rapid way to use the EEG to redirect an imbalanced BAP back into a more normal pattern (12,13) and this reorganizing to a target capability is why the method has direct application to specialized training programs with rigorous selection and acceptance criteria and high costs of operation.

Neural plasticity is the term associated with the brain's ability to reorganize and recover lost functions after injury or illness, and it means that when one part of the brain is damaged or excessively imbalanced, after some re-connecting takes place, the brain's key systems reorganize to a new state of balance where some (or all) of the impaired cognitive or sensorimotor abilities re-emerge. (e.g. a patient regaining the ability to speak or to use his arms and legs again after a stroke). Thus, neural plasticity provides the means for accessing the brain's wiring and directly modulating it to reorganize the activity of its main cognitive and emotional systems to a new state of balance. Even without damage to the brain, it is possible with NT to induce neural plasticity through a process known as operant conditioning, where a stimulus is timed to a particular measure(s) of the participant's BAP and fed back to either reward (reinforce) or punish (extinguish) that particular pattern of brain activity.[14] Several EEG and fMRI studies have reported the use of NT protocols with an ever widening range of notable positive effects correlated with attention, memory, cognitive function and operational performance (15, 16).

3 The Quantitative EEG Normative Database (qEND)

There are a small number of qEND products used clinically for the assessment and treatment of CNS disorders, depression and stress related conditions. Some of these products provide condition-specific medication treatment plans (e.g., Reference EEG), and others are used in assessment, diagnosis and delivery of NT therapies (e.g., NXLink and NeuroGuide). These neuroimaging systems provide an output in the form of color coded maps and graphs that indicate where and by how much the first and second order amplitude and frequency-based features of the EEG differ from an age-matched normal group average; and in which direction the imbalances occur. In clinical care, this information is correlated with other neuropsychological and behavioral assessments and evaluations of the patient, and a treatment plan including neuro-cognitive and cognitive-behavioral strategies is developed to reorganize the patient's BAP and help them develop a positive "way of thinking".

4 Reorganizing Brain Connections: Z-Score Neurofeedback

From a systems perspective, when the brain activity of an individual is out of balance to the point where cognitive and behavioral problems exist, it makes sense to reorganize the connections in key executive and emotional networks to establish a more "normal" BAP. In the most advanced systems, this process uses Z-score guided neurofeedback training (zNT) where real time brain activity is measured, compared to an age-matched norm, and depending on the differences from normal, used to control how a movie is presented to the subject to either reward or punish a particular pattern of activity. For instance, if the delta and theta EEG activity in the frontal lobes of a trainee were to remain below normal, than the picture and sound would be reduced in

clarity and volume as the movie played. Then, when the brain activity moved more towards normal by increasing in delta, theta and even alpha power, the picture and sounds would play more clearly, thereby rewarding the change in his BAP. By repeating this zNT process over several weeks, the trainee's EEG can be guided into a more balanced state with respect to the normal database population used, and in this case, behaviors like impulsivity, anger, anxiety and rumination lessen in severity while mood and cognitive function improve over time (17).

5 Future Assessment, Selection and Neurotraining Pedagogy

In the clinical setting, Z-score guided NT uses software that contains an instantaneous version of the qEND containing all the normalized EEG measures, so a site-specific training protocol can be applied to target the imbalanced brain waves of the patient and move them (through temporal and frequency neuromodulation) in a direction understood to be more normal "or desired" than their current BAP. In practice, it is the average value taken from the "normal population" that becomes the target of the neurofeedback. Then, with repetition, the individual's brain waves can be influenced from an existing "imbalanced" state towards the pre-determined qEND standard BAP. Ever growing research continues to demonstrate that these directed changes in brain activity toward a target pattern of activity are associated with improved task performance, cognitive agility and perceptive functioning, all necessary to achieve persistent resilience to highly stressful situations (18, 19).

With the ability to re-connect and re-organize a trainee's brain waves towards a specific target BAP, it becomes possible to design a qEND from a population of highly experienced and trained mission specialists and then to use that qEND to help select candidates and train them toward a BAP more consistent with the target group normal. With a specifically focused "BAP++ Gold Standard" representing a large group of expert level operators (~500) throughout their career, it may be possible to identify key features of the BAP++ that correlate with reduced stress, increased cognitive agility and elevated motivation and resilience under demanding circumstances. If such a qEND is constructed, then it also becomes possible to design zNT protocols customized for each candidate selected for a specialized program (e.g., Engineer, Pilot, or Operator). In each case, key attributes of the BAP++ would be used to guide the selected trainee's EEG towards a pattern to help them better achieve their training objectives with lower levels of anxiety and negative responses to stress; ultimately leading to higher performance ratings by assessment specialists.

6 Case Study: Research Methods

Two individuals with military leadership experience signed informed consent forms and volunteered to participate in a 20-session Z-score neurofeedback training (zNT) study to evaluate the effects on their BAP as compared to the age-matched means in

the NeuroGuide qEND. These individuals were chosen because of their relevant backgrounds and experience as mission specialists. Subject 1 is a USMC Captain (Res.) with 10+ years in active and reserve service; including deployments in Afghanistan and Haiti and 2 combat tours in Iraq. Subject 2 is a US Army Sergeant Major (Ret.) with over 18 years of service in Special Forces assessment and training.

Immediately prior to and after a 20-session zNT program, 32-channel linked-ear referenced EEG recordings, plus bipolar vertical and horizontal eye, heart and neck muscle channels were recorded during 20-minute eyes-closed resting and 10-minute eyes-open resting conditions. These data were manually reviewed and edited and 2-minutes of non-contaminated EEG data were selected from each subject's EEG for use in the NeuroGuide qEND. The standard 19 channels of the International 10-20 System were submitted for each subject into the NeuroGuide database to produce the Z-scored maps and connectivity graphs displayed in Figures 1 and 2. In each figure, the summary maps on the left are from the recordings made before the intervention and those on the right provide the results immediately after the 20-session zNT trial. The qEND compares the temporal and frequency components of the EEG from all 19 sensor sites (e.g., 1st and 2nd order amplitude and frequency measures from all possible combinations of the 19 sensors) between each subject and the group of age-matched members included in the QEEG database. From those data, it computes the magnitude of the spatial-frequency EEG differences and displays the results in Z-scores, where 1 Z-score is the equivalent of 1 Standard Deviation from the mean for each of the more than 4000 EEG measures computed.

In qEND mapping, the frequency components of the EEG are separated using the FFT, averaged, and displayed in narrow bands: delta (1 – 4 Hz), theta (5 – 8 Hz), alpha (9 – 12 Hz), beta 13- 22 Hz) and high beta (23 – 40 Hz). The magnitude of the difference between a particular EEG measure and the group mean is represented by color coded Z-score maps and graphs, where activity that is -3 std. dev. below normal is shown as Dark Gray and activity +3 std. dev. above normal is Light Gray in color. Abnormal changes in the coherence and phase activity between sensor pairs is indicated by a reduction or excess (Dark Gray lines) as compared to normal, where the thickness of the line indicates the magnitude of the imbalance from +/- 1.96 to 3.09 std. dev. from the mean.

The pilot study evaluated changes in the qEND maps and connectivity graphs (Figs. 1 and 2) from before and after the zNT intervention. Self-reported assessments of anxiety, sleep, depression, job performance and life satisfaction were used to track emotional and behavioral perception. In addition the study examined changes in plasma Cortisol to assess endocrine system stress. To accomplish this, Subject 2 had serial blood draws done throughout his 2.5 week evaluation period: Pre (3 draws), Mid (2 draws) and End (2 draws) to evaluate his corresponding changes in Cortisol level and infer the results to his stress level changes. All seven blood draws were done in the A.M. within 30-minutes each other. Overall, Subject 2's average Cortisol levels went from 14.07 down to 11.45 mcg/dl a reduction of 2.62 mcg/dl (43% from the baseline average, full range = 6.1 mcg/dl).

7 Assessing Neurofeedback Training Results

Thirty-six-channel EEG recordings were made pre and post-intervention from two veterans with leadership experience to track changes in their BAP coincident with a program of 20 zNT sessions (Figures 1 and 2). Their raw EEG signals were reviewed for quality and at least 2-minutes of eyes-open and eyes-closed resting data from throughout the recordings were input to the NeuroGuide qEND to measure individual differences in BAP between each subject and their age-matched group norms.

To design a 4-channel zNT protocol specific to each subject, the results of the behavioral assessments were combined with the qEND results to inform the choice of sites for training. During zNT sessions, an elastic cap was placed on the head that carried the 4 EEG, plus reference and ground sensors, and the wires were attached to the system while the trainee sat comfortably for 30-minutes watching a DVD of their choice. As they watched and listened the trainee's brain waves were measured and used to control how the images of the movie played. The choice of movie was not important, as long as it held the subject's attention.

Figures 1 and 2 each display a pair of eyes-closed EEG Power Summary Reports generated from the Subjects immediately before (Left Report) and after (Right Report) 20 zNT sessions. Visual comparison of the pre-NT and post-NT topographic Power maps (Left vs. Right) details the spatial-frequency differences coincident with the 20 session zNT intervention. Neither of the subjects used medication before or during the evaluation periods and no traumatic events were reported. For Subject 1 (Fig. 1) differences in the Pre vs. Post-intervention Reports were primarily visible in the delta and theta bands where bilateral-frontal and temporal excesses had almost completely resolved, while at the same time the prior reduction of power in the occipital lobes had also normalized (i.e., resolved areas show as Grey in post intervention reports). Subject 2's reports (Fig. 2) indicates a different pattern of change in BAP, where in this case, the delta band shows normalization of only the left occipital lobe imbalance while the right lobe remained in a reduced power state (Note: this may be an indicator of peak performance). The higher frequency bands of beta and high-beta show the largest changes towards normal in Subject 2's BAP, primarily through the reduction of excess power in the right dorsal-medial frontal and insular cortices, all of which play an important role in prosody, empathy, socialization and approach / avoidance behaviors.

Both Subjects conveyed that they experienced improvements in anxiety, rumination, anger, frustration and job performance after 20 zNT sessions. For instance, in the overall assessment, Subject 1's report of anxiety level lowered from an initial score of 49 down to a post-intervention score of 29, a change of 41% over a period of 15 weeks. To evaluate the feasibility of a more rapid training program, Subject 2 carried out two zNT sessions a day and completed the training in 2.5 weeks. He reported a positive change in 5 out of 7 assessments: Insomnia, Depression, Life Satisfaction, and Daytime Function Positive and Negative attributes. In his case, there were no changes reported for the Anxiety or Sleep Quality assessments.

Subject 2 also volunteered to have serial blood draws done to evaluate his morning Cortisol volume changes (measured in mcg/dl). Plasma Cortisol is a biomarker

proportionally related to stress, and provides a quantitative and independent source to monitor changes over the course of the intervention.[20,21] Seven plasma Cortisol measurements were taken at key points over the course of the 2.5 week study. Initial blood draws were taken on the three days prior to the beginning of training to establish a baseline average, two blood draws after the 9th and 11th sessions gave the midpoint average, and two final draws after the 18th and 20th sessions gave the endpoint average. The results indicated a First Half increase in Cortisol from the early to midperiod of 1.28 mcg/dl (a rise of +21%) and a decrease in Cortisol of -3.9 mcg/dl over the Second Half of the intervention (a reduction of 64%). Overall, the average reduction in Cortisol level from beginning to end was 2.6 mcg/dl; a drop of 43% from the baseline. These data are consistent with the subject's reports of feeling less stressed and anxious, and not being so easily angered. This sentiment was also reflected in a relationship assessments filled out by the subject's wife.

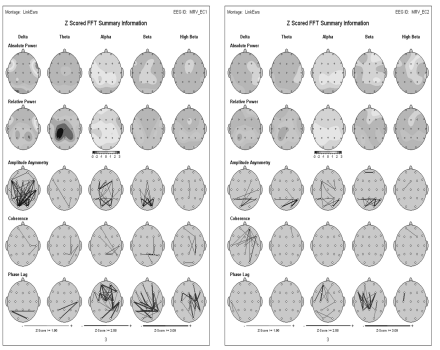


Fig. 1. 33 year old U.S. Marine Captain (Reserve): No diagnosis; high stress, obsessive about home safety. The 2 reports above indicate differences in BAP before (Left) and after 20 zNT sessions (Right) done over a period of 15 weeks. Subject reported lower levels of stress, fewer safety related concerns and improved mood and job satisfaction.

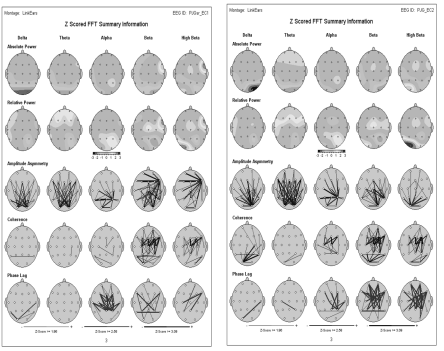


Fig. 2. 44 year old Retired U.S. Army Sgt. Major: No diagnosis; high stress, easily frustrated and angered. The 2 reports above indicate differences in BAP before (left) and after 20 zNT sessions (Right) done over a period of 2.5 weeks. Subject reported less outbursts; lower frustration level and generally improved mood and work performance.

8 Discussion

As we gain a rapidly expanding glimpse into the working brain through a plethora of modern neuroimaging and cognitive-behavioral research tools, we stand at a pinnacle, where brain-machine technologies can externally influence the dynamics and balance of that interconnected system of nervous tissue which constrains the human mind. Even as we stand at the edge, with our minimal understanding of how cells coordinate through waves of chemical and electrical interactions to process information, make decisions and think abstractly, we have made the tools! We are now able and willing to imply meaning to scant measures of mind informing us about the thoughts, moods

and feelings of that person living inside the calvarium…and in good faith we seek to improve the mind within…to make it better when emotions are askew or when one seeks a higher plane of existence or improved levels of performance.

With the infancy of our field beginning to wane, we stand on shaky legs to take those first few steps. Thus we must come together as a community to prevent missteps as we begin in earnest to test the limits of our understanding of mind and machine and develop these impressive new tools in the light of day…not in hiding for fear of public debate. Today it is possible to assess BAP imbalances in one person as compared to a known group of people, and to use the knowledge of these imbalances to redirect specific EEG activity and reorganize the connectivity within that person's brain. This process uses concepts of operant and classical conditioning, targeted neuromodulation and neuroplasticity. There are currently more than 10,000 professional providers of such NT services around the globe with limited regulation and oversight; and only a hand full of government sponsored studies. Yet the adoption of NT by mental and behavioral health professionals and use by consumers has never been at a faster pace. Many NT methods abound supporting a wide range of backgrounds and experience among the providers of clinical services.

Plischke et al.,(22) acknowledge the rapid expansion of neurotechnology businesses and growing exploitation of novice and uninformed users of games, toys, education aides and clinical services. They argue that the use of some neurotechnologies may come with the potential to do harm and call out for regulatory and peer oversight of the brain computer interface industry along with the establishment of a Clinical Field of Practice at the university level. Canli et al., (23) go on to say that it is the responsibility of the entire neuroscience research community to be open about all their endeavors with interfacing neurotechnologies, particularly those related to National Security, and they expect active peer-review and oversight by the researchers themselves, as well as administrators in the governmental funding agencies.

9 Conclusion

Brain waves are changeable towards a predetermined normal pattern of activity using Z-score neurofeedback training and undergoing this process creates a benefit to trainees by lowering their anxiety and stress and helping them better manage their daily behaviors. The zNT approach was successful in two veterans, each with different but relevant backgrounds and military experience. For each subject, the pattern of brain wave activity was different, but the general outcomes from the zNT intervention were the same; reduced stress and improved mood and performance.

Each person handles stress differently, and that difference can be identified in their BAP maps and is likely an inherited trait. Gianotti et al., (24) have identified genetic markers of EEG genotype that link to specific behaviors and they state the purpose of their work saying "it is to identify possible neural mechanisms by which the polymorphism may contribute to stable individual differences. Such neural baseline activation measures are highly heritable and stable overtime, thus an ideal endophenotype

candidate to explain how genes may influence behavior via individual differences in neural function."

Interfacing neurotechnologies are here to stay and the use of normative BAP methods provides the most logical way forward to investigate the use of brain computer systems as a means to reduce stress and improve trainee mental and emotional performance. The question still remains, however, if a customized qEND composed of high performing experts in specialized military domains can be constructed, or if it even should be constructed…but the potential remains, and the feasibility of the success of such a project is very high. Additionally, relevant to military training, it appears possible to compress NT into a shortened time span and still see positive results in stress, mood, life satisfaction and job performance on par with the longer delivery period…borne out in both self-assessments and Cortisol measures. The inexpensive and portable nature of the qEEG method and shortened time frame of a zNT training component makes it feasible as an integrated segment that can be inserted within almost any existing special missions training program.

In the right environment with SMEs experienced in training Special Forces a structured assessment, selection, and training framework combining neurocognitive, metacognitive and cognitive-behavioral methodologies may be more fully investigated. Independently or in combination the use of a BAP++ qEND along with a concomitant real time zNT training methodology provides an approach capable of improving military assessment and training technologies by reducing stress and enhancing cognitive and emotional agility of trainees so they can best process and absorb the tactics and information presented during training.

References

1. Pascual-Leone, A., Freitas, C., Oberman, L., Horvath, J.C., et al.: Characterizing Brain Cortical Plasticity and Network Dynamics Across the Age-Span in Health and Disease with TMS-EEG and TMS-fMRI. Brain Topogr. 24(3-4), 302–315 (2011)
2. Chu, C.J., Kramer, M.A., Pathmanathan, J., Bianchi, M.T., Westover, M.B., Wizon, L., Cash, S.S.: Emergence of stable functional networks in long-term human electroencephalography. J. Neurosci. 32(8), 2703–2713 (2012)
3. Thatcher, R.W., Walker, R.A., Biver, C.J., North, D.M., Curtin, R.: Sensitivity and Specificity of an EEG Normative Database: Validation and Clinical Correlation. J. Neurotherapy 7(3/4), 87–121 (2003)
4. Collura, T.F.: Neuronal Dynamics in Relation to Normative Electroencephalography Assessment and Training. Biofeedback 36(4), 134–139 (2009)
5. Hoedlmoser, K., Pecherstorfer, T., Gruber, G., Anderer, P., Doppelmayr, M., Klimesch, W., Schabus, M.: Instrumental conditioning of human sensorimotor rhythm (12-15 Hz) and its impact on sleep as well as declarative learning. Sleep 31(10), 1401–1408 (2008)
6. Michael, A.J., Krishnaswamy, S., Mohamed, J.: An open label study of the use of EEG biofeedback using beta training to reduce anxiety for patients with cardiac events. Neuropsychiatr. Dis. Treat. 1(4), 357–363 (2005)
7. Giordano, J., DuRousseau, D.R.: Toward Right and Good Use of Brain-Machine Interfacing Neurotechnologies: Ethical Issues, and Implications for Guidelines and Policy. Cog. Technol. 15(2), 5–10 (2011)

8. Quan, M., Zheng, C., Zhang, N., Han, D., Tian, Y., Zhang, T., Yang, Z.: Impairments of behavior, information flow between thalamus and cortex, and prefrontal cortical synaptic plasticity in an animal model of depression. Brain Res. Bull. 85(3-4), 109–116 (2011)
9. Flo, E., Steine, I., Blågstad, T., Grønli, J., Pallesen, S., Portas, C.: Transient changes in frontal alpha asymmetry as a measure of emotional and physical distress during sleep. Brain Res. 1367, 234–249 (2011) (Epub October 1, 2010)
10. Dias-Ferreira, E., Sousa, J.C., Melo, I., Morgado, P., Mesquita, A.R., Cerqueira, J.J., Costa, R.M., Sousa, N.: Chronic stress causes frontostriatal reorganization and affects decision-making. Science 325(5940), 621–625 (2009)
11. Leuchter, A.F., Cook, I.A., Hunter, A.M., Cai, C., Horvath, S.: Resting-State Quantitative Electroencephalography Reveals Increased Neurophysiologic Connectivity in Depression. PLoS ONE 7(2), e32508 (2012), doi:10.1371/journal.pone.0032508.
12. Sirota, A., Buzsáki, G.: Interaction between neocortical and hippocampal networks via slow oscillations. Thalamus Relat. Syst. 3(4), 245–259 (2005)
13. Menon, V., Uddin, L.: Saliency, switching, attention and control: a network model of insula function. Brain Struct. Funct. 214(5-6), 655–667 (2010) (Epub 2010 May 29, 2010), doi:10.1007/s00429-010-0262-0.
14. Brembs, B.: Operant conditioning in invertebrates. Curr. Opin. Neurobiol. 13(6), 710–717 (2003)
15. Gruzelier, J.: A theory of alpha/theta neurofeedback, creative performance enhancement, long distance functional connectivity and psychological integration. Cogn. Process. 10 (suppl. 1), 101–109 (2009) (Epub December 11, 2008)
16. DuRousseau, D.R., Mindlin, G., Insler, J., Levin II: Operational Study to Evaluate Music-Based Neurotraining at Improving Sleep Quality, Mood and Daytime Function in a First Responder Population. Journal Neurotherapy 4, 389–398 (2011)
17. Collura, T.F., Thatcher, R.W.: Clinical benefit to patients suffering from recurrent migraine headaches and who opted to stop medication and take a neurofeedback treatment series. Clin. EEG Neurosci. 42(2), VIII–IX (2011)
18. Scharnowski, F., Hutton, C., Josephs, O., Weiskopf, N., Rees, G.: Improving Visual Perception through Neurofeedback. J. of Neuroscience 32(49), 17830–17841 (2012)
19. Ros, T., Moseley, M.J., Bloom, P.A., Benjamin, L., Parkinson, L.A., Gruzelier, J.H.: Optimizing microsurgical skills with EEG neurofeedback. BMC Neurosci. 10, 87 (2009)
20. Tacker, M.M., Leach, C.S., Owen, C.A., Rummel, J.: Levels of cortisol, corticosterone, cortisone and 11-deoxycoritsol in the plasma of stressed and unstressed subjects. J. Endocrinol. 76(1), 165–166 (1978)
21. Swigar, M.E., Kolakowska, T., Quinlan, D.: Plasma cortisol levels in depression and other psychiatric disorders: a study of newly admitted psychiatric patients. Psychol. Med. 9(3), 449–455 (1979)
22. Plischke, H., DuRousseau, D., Giordano, J.: EEG-based Neurofeedback– The Promise of Neurotechnology and Need for Neuroethically-informed Guidelines and Policies. J. Ethics Biol. Engineer. Med. (July 2012), doi:10.1615/EthicsBiologyEngMed.2012004853
23. Canli, T., Brandon, S., Casebeer, W., Crowley, P.J., DuRousseau, D., Greely, H., Güzeldere, G., Pascual-Leone, A.: Neuroethics and National Security. The American Journal of Bioethics 7(5), 3–13 (2007)
24. Gianotti, L.R.R., Figner, B., Ebstein, R.P., Knoch, D.: Why some people discount more than others: baseline activation in the dorsal PFC mediates the link between COMT genotype and impatient choice. Frontiers in Neuroscience, Decision Neuroscience 6, Article 54, 1–12 (2012)

Ecological Momentary Storytelling: Bringing Down Organizational Stress through Qualifying Work Life Stories

Lisbeth Højbjerg Kappelgaard[1] and Katja Lund[2]

[1] Centre for Dialogue and Organisation, Department of Communication and Psychology, Aalborg University, Denmark
lisbethhk@hum.aau.dk
[2] Interactive Digital Media, Department of Communication and Psychology, Aalborg University, Denmark
katja@hum.aau.dk

Abstract. The purpose of this article is to examine ways in which a combination of ecological momentary assessments and reflective dialogues can provide a methodological framework for qualifying work-life stories in the process of reducing organizational stress. The article is based on two hypotheses: 1) a general as well as a work-related sense of coherence can mobilize resistance to stressors and 2) a sense of coherence can occur through self-reflective narratives which clarify patterns of action for oneself and for others. Focusing on hearing impaired people in the Danish work force as well as primary school teachers, the authors create a stress tracking method based on HRV-measurements coupled with mobile questionnaires and reflective dialogues. Findings in the user-test indicated that the method is a tool that creates a story-based foundation on which it is possible to start a process of talking about own experiences, stress and stressors, strategies, contexts etc. when dealing with organizational stress.

Keywords: Ecological Momentary Assessments (EMA), organizational stress, Experience Sampling Method (ESM), Heart Rate Variability (HRV), Sense of Coherence (SOC).

1 Introduction

Research shows that an increasing number of people in the working population suffer from occupational stress. This is illustrated by the fact that several occupational medicine clinics in Denmark have experienced twice as many referrals of patients over the past five years [9]. Several reports conclude that there is a need for research whose results can immediately be converted and used in the practical efforts of organizations to improve the working environment. Also, interdisciplinary and solution-oriented research based on a holistic approach and the involvement of users in research, planning, implementation and dissemination have been called for [2].

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 572–581, 2013.

1.1 Background

Since 2011, both authors have conducted parallel research in the field of uncovering the growing problem of work-related stress under the auspices of Aalborg University (AAU), Department of Communication and Psychology. One project has a special focus on communication and stress among hearing impaired people in the Danish work force, while the focal point of the second project is to identify which discourses on work-related stress are produced in the professional field of teachers.

In addition to work-related stress as a common target field, the authors shared a methodical ambition as both projects aimed to reflect a holistic perception of stress. Both projects adhered to the basic assumption that stress is a term with several, fundamentally inseparable, dimensions, and both authors worked with a holistic, interdisciplinary, bio-psycho-social stress concept [11]. In the wake of this understanding of stress, the authors also assumed that in the attempt to elucidate stress, it is necessary to focus on spoken as well as tacit knowledge; spoken knowledge meaning knowledge which is linguistic, rational and articulated, and tacit knowledge meaning knowledge that is not immediately articulated and accessible, yet producing meaningful stories such as bodily experience, memory and behaviour.

From this theoretical starting point followed methodical frustration: how was it possible to reflect a holistic understanding of stress which focuses, at the same time, on body, mind and social factors?

1.2 When "Quantify Yourself" Paves the Way for "Qualify Yourself"

In November 2011, both authors participated in the pilot project "Quantify Yourself". This was a methodical turning point. The test lasted a week and was conducted under the research unit "Humansensing", AAU, where four types of Ecological Momentary Assessments (EMA) were tested in combination with each other: Heart Rate Variability (HRV), Galvanic Skin Response (GSR), GPS and an online questionnaire to be answered once an hour. Subjects carried the HRV and GSR sensors as well as a GPS around the clock and accessed the questionnaire with a smartphone.

The authors felt on our own bodies how the combination of different EMA sources provided increased awareness of how, in our daily activities, we manage our energy and react physically and mentally to specific situations and activities. This led us to address questions such as: *Is there anything I should do differently in my life? Why do I sometimes act in a way that basically does not seem to work for me - which brings me mentally or physically to my knees? Why don't I do more of what seems to contribute positively to me - that provides energy?* The responses became a form of electronic diary or a life story that helped us to mirror ourselves and retain memories of behaviour patterns.

At the same time, we felt the value of seeing the spoken and tacit knowledge in a context. The responses to the questionnaires were our spoken knowledge - we could reflect on the experienced energy level and mood and express it. The physical measurements were basically tacit knowledge - we did not have access to knowledge about the moisture level in our skin and why it would increase or decrease. Nor did

we have access to knowledge of our exact heart rate or HRV - but access to the tacit knowledge in combination with the spoken knowledge would provide valuable knowledge through stories about behaviour patterns and the management of energy level.

During the pilot project, both authors also experienced the lack of collective, reflective space where we could speak these obtained stories out loud and retain the knowledge we had gained in the past week. Out of this personal experience grew the idea that some adjustment of the method may help to create the reflective space that can qualify work life stories, thus offering an opportunity to create an increased sense of coherence among people who are particularly vulnerable to experiencing stress at work. Our basic assumption is that a method development based on different approaches to EMA combined with reflective dialogues will contribute constructively to articulating the experience of work-related stress and thus open up for increased action at an individual as well as a group level.

2 Research Question

How can a methodical combination of Ecological Momentary Assessments qualify work-life stories that can provide greater insight into and understanding of work-related stress?

3 Theoretical Foundations

The theoretical inspiration for the development of a stress tracking method was taken from the linking of Ecological Momentary Assessment, medical sociology and humanistic psychology. This frame is explicated below.

3.1 Ecological Momentary Assessment

Ecological Momentary Assessment (EMA) is a term that covers a wide range of research methods and traditions, all of which have in common that they provide access to data on the subject's movements in the present and in the specific environment. Examples of EMA may vary from traditional diary keeping to the collection of biosensor-data and online activity logs. Thus, there are various categories of EMA: “Experience Sampling Method” (ESM) is registered subjectively experienced states; “self-monitoring” is records of actions; “ambulatory monitoring” detects the subject's physiological state [10].

In this article, we are dealing with a combination of ESM in the shape of a mobile questionnaire and ambulatory monitoring represented by HRV biosensor-data.

Experience Sampling Method. The method is particularly suitable for gaining an insight into social, psychological and physiological processes and experiences in the present [3]. It is the spontaneous here-and-now response that is captured, thus

avoiding the biases that might be associated with reflective and memory-based data acquisition [10]. Memory can be selective and has often, in qualitative research as in the treatment of stress methods, been based on interviews and dialogues about experiences that might be months or even years old. The authors acknowledge the value and significance of such stories but also hold the basic assumption that a different time perspective closer to the moment when an experience occurs might offer a different picture of an incident, an experience, a feeling etc. Building the method on EMA combined with a qualitative approach is an attempt to embrace and accommodate both long-term and short-term stories.

Ambulatory Monitoring. The biofeedback gives us access to the tacit knowledge produced by the physical body [4, 6]. We feel our bodies react when we start to sweat or when the heart is pounding when faced with a challenging situation. But we might not notice small differences in the body's signals that might give away feelings of mental distress or experiences of stress. In continuation of our holistic stress understanding, it is a basic assumption that we cannot always rationalize our way to understanding. Body and mind must be reconciled. If we isolate the action from the body and exclusively connect it to the mind, we ignore the essential human condition that the self is a unity of body and mind [5].

3.2 Sense of Coherence

A hypothesis in the method development is that the sense of coherence creates a resistance to stressors - a hypothesis derived from medical sociology. We are inspired by Aaron Antonovsky's salutogenetic idea [1] which, in stead of focusing on that which leads to disease (pathogenesis), focuses on that which leads to health and resistance to disease (salutogenesis). Antonovsky's premise is that throughout life, all people are affected by a varying number of stressors. Antonovsky was particularly interested in investigating how people mobilize a resistance to the stressors. What determines how an individual manages to get on with his or her life when challenged with great resistance? According to Antonovsky, the answer to this question is the concept of sense of coherence (SOC) [1]. The main point is that the better we are able to see the coherence of different contexts in our lives, the greater the resistance we are able to mobilize against the stressors that life offers us [1].

According to Antonovsky, SOC represents a life-long learning curve. This learning process has the best conditions when we are experiencing life as comprehensible, manageable and meaningful, which are the three key components of the concept of SOC [1].

3.3 Recognition through Dialogue

Our understanding of the concept of dialogue is based on humanistic psychology. We are particularly inspired by Kristiansen and Bloch-Poulsen [7] who define dialogue as

unpredictable, risky and exploratory conversations where truth is not predetermined but where recognition is produced in the interpersonal contact. The aim is to jointly produce new insights or options. Central to this dialogue understanding is that dialogue is not only skills but also an interpersonal way of acting towards each other – it is a way of being. In this regard, we are particularly inspired by Carl Rogers' 3 concepts: congruence, empathy and affirmation, which in our view is crucial to be present with and for the other in the dialogue [8].

3.4 EMA as a Creator of Unifying Work-Life Stories

"If we keep our eyes wide open to reality, the way is open to an increasing understanding", Antonovsky writes [1]. At the same time he writes that the way to understanding and to "opening one's eyes" goes through a person's life stories - this is where the meaningful may occur. Furthermore, writes Antonovsky, it is not certain that a person with a strong SOC has a plan of action. Thus, a person may well feel paralyzed or miss the reflective space that can put them in a position to act constructively. We argue as our second hypothesis that stories and reflections pave the way for action and change, and that these stories and reflections are to be captured and articulated through a methodical combination of EMA and reflective dialogues.

4 Method Framing

The authors conducted the pilot test that was constructed in order to get feedback on both the functionality of the system as well as the method and the structuring of the content and questions. In this article we will not dwell on feedback on the system but rather on the method as a whole.

We chose to test the method on a person representing each of the two groups the authors work with on a daily basis namely teachers and people with a hearing loss.

Questionnaires. In view of the on-going practice studies for both user groups, the ESM was designed as a mobile questionnaire. The aim is to develop an application that is generally applicable.

Based on general knowledge on stress and specific knowledge on the two user groups, what we finally wished to acquire through the questionnaire was information on: 1) situation (activity the person is involved in and how many people are in the same room), 2) energy level and mood, 3) SOC (the three components comprehensibility, manageability and meaningfulness).

At the end of the questionnaire, there is the opportunity to write additional text, take a photo or record audio either to measure the noise level or to elaborate on the situation and add thoughts of the moment.

User Test. The user test was based on 3 main modules:

Day	Activity
1	• The test persons are introduced to the project • A mobile phone with the application is handed out • The test persons are introduced to the use of the system (questionnaire and HRV equipment), and the HRV equipment is switched on and attached to the chest • The test persons starts filling in the questionnaire once an hour
2 - 4	Testing
5	A follow-up dialogue based on a reflection exercise and data analysis (day 4 activities) is implemented

5 User Test Findings

Through user testing we sought to answer the question:

How can a methodical combination of Ecological Momentary Assessments qualify work-life stories that can provide greater insight into and understanding of work-related stress?

The following findings from the test run aim at answering the above.

Can a Sense of Coherence Be Tracked? One of the hypotheses in the development of the method is that the sense of coherence (SOC) can mobilize resistance to stressors. Therefore, during the test process, we were particularly interested in whether or not, through the method, the test person had the experience of finding greater coherence or a space for greater reflection on the contexts in his life.

5.1 Findings – Test Person 1 (TP1)

The following are statements from the follow-up dialogue with TP1:

"It amazes me that I had so much energy when I worked in the evenings throughout the week leading up to the deadline Friday. On the other hand, I was completely exhausted Saturday. I have not thought about it much before how demanding it is and how exhausted I am physically and mentally after such a deadline has been reached." and Our interpretation of these statements is that through self-monitoring and responding to ESM, TP1 obtains a meta perspective on his own practices and ways of managing energy. The data analysis supplies TP1 with a new insight on both a mental and a physical level, and it provides an elevated sense of coherence through a greater understanding of how different elements that constitute one's life are connected and affect one's actions, reactions and behaviours. It becomes obvious what price his body and mind pay after having reached a deadline. The test person said that he had not previously reflected on how much he subsequently responds to such pressure.

During the test period, TP1 had the – in his own words - *"…privilege only to have to focus on finishing the paper"*. In the follow-up dialogue it became clear that the focus on only one task had given him an extra amount of energy to complete the task and he felt an elevated sense of meaning. To the question: *"How can you use this information?"* he replied, *"I can use the information to see how important it is for me to have a meaningful task. It makes sense to me - it gives me energy, whereas tasks that do not make sense steal my energy."*

In particular, self-monitoring of HRV appears to be meaningful to TP1. His first comment during feedback on this was:

"Everyone should have access to this! I map many of the activities I do - use my calendar a lot. Here it is interesting for me to see how my body reacts for example when I work in the evening, go for a walk in the city or after an important deadline. Not everyone uses a calendar the way I do and for those who don't I think that this type of questionnaire is a great way to remember what you have been up to… ".

Overall, the statements indicate that the method can be used as a reflexive space that can qualify work-life stories. But several statements also indicated that the introduction and follow-up dialogues with a test leader who can help explain and clarify concepts and data analysis as well as induce a larger degree of reflection are necessary. For example, during the follow-up dialogue TP1 reflected further on the concepts of balance, overview and meaning:

"I still find it difficult to interpret the concepts balance, overview and meaning. I find myself reflecting on what you mean by this? "

Several times throughout the follow-up dialogue, questions about the meaning of concepts are asked. This draws our attention to the importance of giving a thorough initial introduction where the concepts are discussed and explained on the basis of the understanding and the situation of the test person.

5.2 Findings Test Person 2 (TP2)

TP2 sums up his experience of being involved in the test as follows:

"In general I can say that I have been confirmed in the feeling I had that my mood is often very positive. Moreover, I think that the contexts and people I surround myself with during a normal workday as well as in my spare time have a positive influence on my mental and physical balance".

This shows the method to be useful in discovering patterns and contexts that can explain certain feelings or a certain level of SOC as it clarifies connections between internal and external factors.

When going through the HRV analysis, the dialogue becomes particularly relevant as it can be difficult to see the difference between a situation of physical activity and a stressful situation. TP2 was interested in knowing how his HRV was affected in a specific conflict situation, and the test leader analysed the time of the conflict on the

HRV measurement. It was obvious that he was emotionally and physically affected as he explained that he had to stop a conflict between two of the pupils. His HRV at that time was almost identical to the measurements a moment before when he had been carrying a heavy box up the stairs.

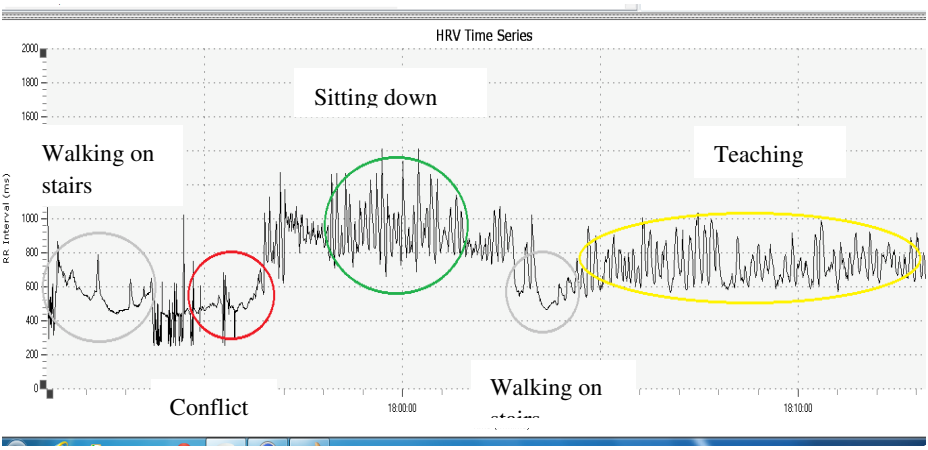


Fig. 1. The x-axis illustrates the R-R interval, which is the period of a heartbeat measured in milliseconds. Increases in stress are associated with decreases in RR interval.[1]

Identifying a moment like that is of great importance as it gives the test person a possibility to connect the feeling of distress he experienced at that moment with his bodily experience and from that point reflect on ways in which to deal with future conflict situations. The dialogue went on to reflect on his ability to calm down immediately after a stressful situation, which seemed to be comforting him. Also the fact that his HRV was fairly stable during lessons when having to correct the pupils once in a while seemed to have a positive effect on his perception of self. The data from the HRV was compared to the ESM data, which showed that his energy level that day dropped gradually. This had also happened the day before, however, and as we only had 4 days of measurements, it could be a coincidence rather than the expression of a certain connection. TP2 expressed several times that he would have preferred a longer test period in search of more significant patterns.

TP2 reflects on the role of the test leader when talking about the possibility of taking more long-term measurements and analysing data him-self. He says: *"I do not think one should underestimate the value of another person who has not been in the situation to analyse the numbers - looking at it from a different angle might generate different reflections"*. Furthermore to the question: *"Has it made sense to you to be involved in the testing process?"* TP2 responds: *"Yes, but it only really makes sense when sitting down talking about it."*

[1] The y-axis is the time line. In this example the HRV–equipment was out of sync with the actual time as this HRV-measurement was started at 9.59 am.

In order to become aware of the reasons for engaging in different activities TP2 says: *"... what you write*[2] *about me scoring high in both energy and mood when I'm engaged in sports activities… these are actually some girls that I coach, and one could ask: why do you coach some girls you don't even know? Well that's because it makes me happy."*

The reflection above underlines the method's ability to find connections and reflect more deeply upon choices in one's life.

Finally he sees the test system as a potentially very useful tool for working in teacher teams, particularly with new school structures in mind, where several teachers and classes merge together for longer or shorter periods of time.

6 Conclusions and Reflections

Meta Perspective as an Impetus for the Creation of Work Life Stories. The main findings of this process indicate that the method is useful in creating the space for reflection and work life stories. Both test persons 1 and 2 express that the ESM and HRV monitoring provided them with a meta perspective on the extent to which their daily activities impact them both physically and mentally. In the follow-up dialogues, both test persons articulate how the method can serve as a useful step for them to gaining more insight into the connection between situations and activities in their work lives and their wellbeing in general.

Need for Methodological Triangulation. The test also proved each part of data acquisition to be essential to the method in order for the test persons to reach a better understanding of their own work life stories and in order for the authors to achieve an accurate understanding of the data. It became clear to the authors that the ESM, the bio-feedback or the dialogues are not sufficient separately. The authors experienced how crucial the dialogues are in order to understand the HRV-data. The authors could have interpreted the HRV curves in many ways, but we could not reach an accurate understanding without the follow-up dialogues. On the other hand, dialogues without ESM and HRV-monitoring would reduce the authors' and the test persons' insight into both bodily and mental here-and-now reactions.

Though we have reason to believe that this method can qualify work life stories and bring us closer to an understanding of work-related stress, the test also prompted the authors to reflect on some of the potential biases and pitfalls that this method may contain.

An Individualizing Trap? The answer to the salutogenetic question is the individual's sense of coherence, which is a lifelong learning process to which the individual himself may open the door through reflections and the creating of life stories. An objection to this particular standpoint might be that on this very point, the method could contribute to the individualization of the stress problem. In this context, individualization signifies the risk that the responsibility for the extent to which the individual copes with everyday stressors may become the individual's own business.

[2] Both test persons received written feedback with an analysis of the ESM and HRV data before the follow-up dialogue.

The individual "owns" the learning process. It may therefore be argued that the responsibility for whether or not the individual is able to act constructively in various situations rests solely on the individual. Phrased differently: If things cease to make sense or lack the sense of coherence, this is the individual's own fault! The question is whether, on the basis of Antonovsky's concepts, we are moving towards a scenario where SOC applies to the individual alone and therefore undermines the articulation of critical conditions in the individual's work life – conditions which are beyond the individual's responsibility. On this matter, it is very important for the authors to emphasize that SOC cannot be reduced to a psychological characteristic that directs behavior. It is our belief that SOC – or the lack of it - occurs in a dialectical relationship between the individual and his/her surroundings.

Ethical Issues Regarding the Data. In this regard, the authors also find it crucial to consider ethical issues pertaining to the data. An organizational context is an arena with many interests and relations which are both symmetric and asymmetric. Power is at stake between employees and managers. Goals differ. Therefore, the authors emphasize that this method in itself should be used with great care and include reflection on at least the following questions: Who owns the data? Who gains an insight into what and why? What purpose does the monitoring serve?

References

1. Antonovsky, A.: Helbredets mysterium, pp. 33–45. Hans Reitzels Forlag (2000)
2. Arbejdsmiljøforskningsfonden/COWI: Kortlægning og analyse af dansk arbejdsmiljøforskning. Ministry of labour (2006)
3. Beal, D.J., Weiss, H.M.: Methods of Ecological Momentary Assessment in organizational research. Organ. Res. Methods 6, 440–464 (2003)
4. Eller, N.H., Kristiansen, J., Hansen, Å.M.: Long-term effects of psychosocial factors of home and work on biomarkers of stress. Int. J. Pscyhophysiol. 79(2), 195–202 (2011)
5. Hastrup, K.: Inkorporeret viden og praktisk kunnen, i Red. In: Baarts, C., Fredslund, H. (eds.) Perspektivet – kvalitativ forskning i arbejdsmiljø og arbejdsliv, pp. 45–47. Arbejdsmiljøinstituttet (2005)
6. Kristiansen, J., Mathiesen, L., Nielsen, P.K., Hansen, Å.M., Shibuya, H., Petersen, H.M., Lund, S.P., Skotte, J., Jørgensen, M.B., Søgaard, K.: Stress reactions to cognitively demanding tasks and open-plan office noise. Int. Arch. Occup. Environ. Health 82, 631–641 (2009), http://www.ncbi.nlm.nih.gov/pubmed/18936956
7. Kristiansen, M., Bloch-Poulsen, J.: Kærlig rummelighed i dialoger – om interpersonel organisationskommunikation, Institut for Kommunikation, pp. 15–16, Aalborg Universitetsforlag, 1. udg. 1. oplag (2000)
8. Rogers, C.R.: The interpersonal relationship. The core of guidance. Harvard Educational Review 32(4) (1962)
9. Rohde, B.: Stor stigning i antallet af arbejdsmedicinske patienter, i Arbejdsmiljø, 4
10. Shiffman, S., Stone, A., Hufford, M.: Ecological Momentary Assessment. Annu. Rev. Clin. Psychol. 4, 1–32 (2008)
11. Zachariae, B.: Stress i et biopsykosocialt perspektiv.I Red. In: Damsgaard-Sørensen, K., Madsen, B. (eds.) Stress – når kroppen siger fra, pp. 14–15. Gyldendals Akademiske bogklub (2003)

The Development and Application of a Novel Physiological Metric of Cognitive Workload

Jeremy C. Rietschel[1] and Matthew W. Miller[2]

[1] Maryland Exercise and Robotics Center of Excellence
VA Rehabilitation Research & Development
Baltimore, MD 21201, USA
jcrietschel@yahoo.com
[2] Department of Kinesiology
Auburn University
Auburn, AL 36849, USA
mwm0024@auburn.edu

Abstract. An objective assessment of the cognitive burden imposed by a task (cognitive workload) is of fundamental interest in that it would provide a "window" into one's current allocation of cognitive resources. Such insight would have tremendous implications in maximizing human performance through a multitude of applications including human-computer interaction. The authors propose a novel, electroencephalographic (EEG)-derived metric, which relies on the event-related potential (ERP) component, novelty-P3. A theoretical rationale and experimental evidence supporting the metric's utility are provided, followed by future directions.

Keywords: Cognitive workload, human performance, EEG, novelty-P3.

1 Introduction

This paper will present a novel method to assess the cognitive burden imposed when one performs a task (i.e., cognitive workload). First, the importance of such a metric will be discussed followed by how this metric was conceptualized and developed. Next, three experiments aimed at validating the capability of this measure regarding the assessment of cognitive workload will be presented. Finally, the paper will conclude with recommendations for future research regarding this metric.

1.1 Why Measure Cognitive Workload?

An accurate measure of cognitive workload would be useful in a multitude of ways. For instance, one would be able to determine how different task conditions impact the mental state. This information would be useful in designing a task so as to reduce excessive cognitive workload and limit mental fatigue. Additionally, a cognitive workload assessment could serve as a forecast of future behavior. For example, two

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 582–590, 2013.

individuals could be executing the same task at comparable levels of performance and thus would be indistinguishable from each other using a behavioral level of analysis. However, it could be that one individual is performing the task at a considerable cognitive 'cost' whereas the other individual is able to perform similarly with little strain placed on his/her cognitive resources. Knowing this, one could predict which individual could maintain his/her level of performance longer and who would be better able to cope with unexpected increases in task demands (the latter individual). Similarly, task mastery has been robustly associated with automaticity (the ability to perform a task with little mental effort), and as such, measuring cognitive workload would inform skill level beyond that of looking at the performance alone. In addition, continual monitoring of cognitive workload would reveal the dynamic mental state of an individual. This information could be used to maximize user/task interaction by adjusting task demands to match the user's current cognitive state. For example, if a cognitive workload metric detects that an aircraft pilot is experiencing excessive cognitive workload while flying the aircraft, the machine (aircraft) could assume task demands by engaging an autopilot feature. Similarly, in a team environment task, demands could be dynamically allocated among team members based on their respective cognitive workloads such that each member maintains a manageable load. For pictorial examples illustrating the utility of a metric assessing cognitive workload, see Figure 1.

1.2 Background and Development

Cognitive resources are limited in regards to quantity [1]. As one engages in a task, the cognitive workload imposed by the task draws upon these finite cognitive resources. The spare resources not currently being utilized by the task are referred to as attentional reserve and are available to allocate to additional task demands (e.g., unexpected events). In this regard, cognitive workload and attentional reserve are inversely related such that when cognitive workload increases, attentional reserve decreases. Conversely, when cognitive workload is reduced, attentional reserve grows [2]; see Figure 2A. Therefore, assessing attentional reserve provides insight into the current state of cognitive workload.

Thus, in order to develop a technique to measure cognitive workload, we sought to objectively quantify attentional reserve using a neurobiological approach. The electroencephalographic (EEG) technique measures the electrical activity of the brain. Brain activity associated with the processing of stimuli can be assessed by extracting a portion of the EEG signal time-locked to the onset of the stimuli—these EEG segments are known as event-related potentials (ERPs). ERPs are comprised of different components, each of which reflects a distinct cognitive process. The component known as the novelty-P3 reflects the automatic orienting of attention to novel stimuli, and the amplitude of the novelty-P3 component is positively related to the degree of this cognitive process [3]. The degree to which attention can be oriented to novel stimuli depends on the availability of cognitive resources for such orienting (i.e., the magnitude of attentional reserve). Thus, we reasoned that novelty-P3 component amplitude would reflect the quantity of attentional reserve. Specifically, when attentional

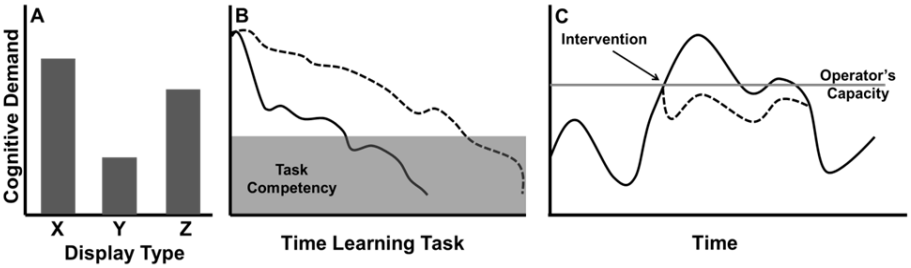


Fig. 1. Illustrative scenarios demonstrating the utility of an assessment of cognitive workload. A) In order to determine the most efficient means to convey information, three different visual displays representing the same information are presented to individuals. Their cognitive workload is assessed during these presentations and it is revealed that display Y conveys the information with the least cognitive demand imposed, thus it is the most efficient. B) In order to determine when a flight controller has had enough training to begin real-world operation their cognitive workload is assessed as they learn how to perform their task. When they can perform the task with minimal cognitive workload (i.e., perform the task below a specified threshold of cognitive demand—"task competency"), then they are considered adequately trained. In the current example the two tracings correspond to two trainees with the solid line representing a trainee who reached the competency threshold quicker than the trainee represented by the dotted line. C) During the dynamic production of a task, if the cognitive demand associated with the task exceeds the operator's capacity then the probability of failure greatly increases. The ability to monitor the cognitive workload during task production (solid line) would inform when demand is exceeding capacity, which could trigger an intervention aimed at reducing the demand (dotted line) thus averting the increased risk of failure. For example, if a pilot became overloaded during a flight, then the co-pilot could begin to take over some of the responsibilities, effectively reducing the pilot's load and, thus, the probability of an accident.

reserve is high, many cognitive resources are available to be oriented to novel stimuli, which should then be reflected by large novelty-P3 component amplitudes. Conversely, when attentional reserve is lower, fewer cognitive resources are available to be oriented to novel stimuli, which should result in reduced amplitude (see Figure 2B). Given the inverse relationship between attentional reserve and cognitive workload, we predicted high cognitive workload should result in small novelty-P3 amplitude, whereas lower cognitive workload should result in larger novelty-P3 amplitude. In this regard, we predicted the novelty-P3 component should be effective in assessing cognitive workload.

In line with this rationale, our approach in assessing cognitive workload involves probing individuals with stimuli known to elicit the novelty-P3 component while they engage in a primary task (a task for which cognitive workload measurement is of interest). Specifically, we present individuals with novel, task-irrelevant, ecologically-valid auditory stimuli (e.g., a woman coughing, a dog barking, a glass breaking). Concurrently, EEG is recorded and time-locked to the stimuli. Next, ERPs to the stimuli are extracted and the average amplitude of the novelty-P3 is computed.

There are three distinct advantages to this approach. First, the EEG signal is an objective assessment and thus not influenced by the subjectivity typically introduced when employing self-report methods of cognitive workload assessment. Second, the

most commonly employed method used to measure cognitive workload (i.e., the dual-task paradigm; [e.g., 4-7]) may risk inherently confounding the assessment [8-9]. In dual-task paradigms, participants are probed with stimuli to which they are asked to attend (secondary task) while performing the primary task. For example, participants may be asked to count auditory stimuli (secondary task) while performing a simulated aircraft flight (primary task). The major limitation of such paradigms is that the addition of a having to attend to secondary task stimuli may fundamentally interact with the primary task, thus compromising the magnitude of cognitive workload imposed by the primary task alone. As our method probes individuals with task-irrelevant stimuli (i.e., stimuli to which individuals are not instructed to attend), it avoids this limitation altogether. Third, we probe individuals with ecologically-valid, novel stimuli. The salience of such stimuli is believed to induce a compulsory orienting of spare cognitive resources [3]. Therefore, this method is likely to provide a robust assessment of attentional reserve and thus cognitive workload.

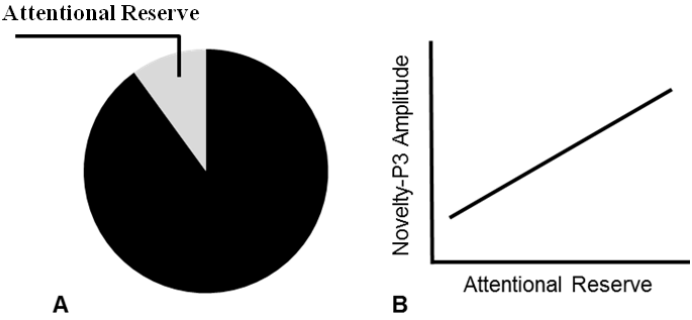


Fig. 2. A) The conceptual model indicating (1) that cognitive resources are fixed with regard to total capacity, and (2) when a cognitive workload is imposed, the resources that are spared are referred to as attentional reserve. Accordingly, this relationship reveals that measuring attentional reserve will, in turn, reveal the magnitude of cognitive workload. B) Hypothesized relationship between attentional reserve and novelty-P3 amplitude. As attentional reserve increases, this is reflected in increased novelty-P3 amplitude. Conversely, as attentional reserve decreases, novelty-P3 amplitude becomes reduced.

2 Experimental Assessment of the Metric

2.1 Experiment 1

The first experiment aimed at testing the validity of our cognitive workload metric involved incrementally varying the difficulty of a primary task [10]. We reasoned that increasing task difficulty would elicit a corresponding cognitive burden, thus raising cognitive workload. Therefore, we predicted that incremental modulations in task difficulty would induce dose-dependent changes in cognitive workload and, as such, our metric should be sensitive to these changes.

Twenty participants performed the videogame Tetris at three levels of difficulty presented in random order: View, Easy, and Hard. Tetris requires individuals to use a keyboard to manipulate different-shaped game pieces presented on a video screen in order to place them in an optimal location (limiting the space between the current piece's placement and previously played pieces). During the View level of difficulty, participants watched Tetris but did not manipulate the game pieces. This level was expected to impose the least cognitive burden as individuals did not directly interact with the game. During the Easy level, participants maneuvered game pieces moving down the video screen at a velocity of 1.67 cm/s, whereas during the Hard level participants manipulated pieces moving at 3.56 cm/s. This difference in speed was believed to elicit greater cognitive workload in the Hard condition as compared to the Easy, as participants had to more quickly decide where to place the current game piece, execute the placement, and update their planning for successive pieces. During each level, we employed our cognitive workload assessment. Specifically, participants were probed with novel, task-irrelevant, ecologically-valid auditory stimuli. Concurrently, EEG was recorded and time-locked to the stimuli. Next, ERPs to the stimuli were extracted and the average amplitude of the novelty-P3 was computed.

Behavioral results revealed poorer task performance in the Hard level than the Easy level, suggesting a successful manipulation of task difficulty. As predicted, novelty-P3 amplitude incrementally changed as a function of task difficulty level. Specifically, novelty-P3 amplitude was largest in the View level, second-largest in the Easy level, and smallest in the Hard level (see Figure 3). These results suggest that our metric is able to provide an effective assessment of cognitive workload. Specifically, a negative relationship between cognitive workload and novelty-P3 amplitude was observed, which is consistent with our conceptual model.

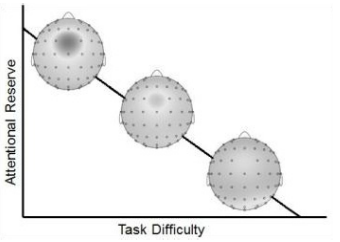


Fig. 3. Support of our conceptual model. Specifically, as a task becomes more difficult, cognitive workload increases, resulting in reduced attentional reserve. The scalp maps of the novelty-P3 are actual data from the three difficulty levels: View, Easy, and Hard (darker grey indicates higher novelty-P3 amplitude). As expected, novelty-P3 amplitude was inversely related to task difficulty, suggesting that our metric is able to provide an effective assessment of cognitive workload.

2.2 Experiment 2

In our second experiment task difficulty was held constant while participants' skill level improved [11]. It is generally accepted that as individuals learn a new task, the

cognitive workload required to perform that task becomes reduced [12]. Accordingly, we sought to examine if our metric was sensitive to changes in cognitive workload related to individuals' current skill level.

Twenty-one participants all performed a center-out reaching task that required moving as quickly and accurately as possible to targets. However, they were randomly assigned to either a group that learned a novel visuomotor distortion (i.e., requires learning) or to a control group that performed the same task with no distortion element (i.e., no learning). For the duration of the task, our metric was employed to assess cognitive workload. We predicted novelty-P3 amplitude would initially be low in the Learning group relative to the Control group, but that there would be a progressive increase in amplitude in the Learning group as a function of learning. Additionally, as the Control group was not required to learn the distortion, we predicted that novelty-P3 amplitude would remain relatively stable.

Behavioral evidence supported that the Learning group experienced learning whereas the control group did not (i.e., the Learning group significantly improved task performance, whereas the Control group's performance remained stable). As expected, the Learning group exhibited a progressive increase in novelty-P3 amplitude over the course of learning, whereas the Control group did not exhibit significant changes in amplitude (Figure 4). In other words, across the time period where individuals learned a new skill, our metric revealed a progressive decrease in their cognitive workload. Moreover, our metric revealed no change in the cognitive workload of individuals assessed across the same time period, performing the same task but without the learning component. Thus, our metric was sensitive to predictable changes in cognitive workload associated with skill learning.

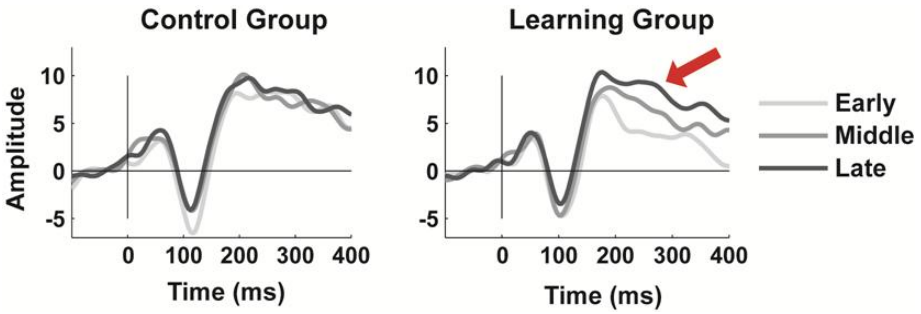


Fig. 4. The change in novelty-P3 amplitude as a function of skill level. On the right panel, the Learning group's ERPs to the auditory stimuli are presented. The light grey, darker grey, and black lines correspond to early, middle, and late learning, respectively. Consistent with predictions, novelty-P3 amplitude (indicated by the arrow) become larger as a function of learning, supporting our method as a valid assessment of cognitive workload. On the left panel, the Control group's data are presented. As expected, there was no change in novelty-P3 amplitude over the course of task performance.

2.3 Experiment 3

In our third experiment, task difficulty and participants' skill levels were held constant while participants' environments were manipulated [13]. Specifically, 12 participants performed Tetris at a difficulty level yoked to his/her respective skill level in two social environments: a high quality team environment and a low quality team environment. In the High Quality Team Environment, participants performed Tetris with a teammate who they perceived as being competent. Conversely, in the Low Quality Team Environment, participants performed with a teammate who they perceived as being incompetent. Prior research has indicated that individuals performing in high quality team environments experience significantly reduced cognitive workload relative to performing in lower quality environments [14]. Accordingly we sought to examine if our metric was sensitive to changes in cognitive workload related to this aspect of the social environment.

Participants reported, via a questionnaire, that the High Quality Team Environment was the preferred social environment. As expected, participants exhibited higher novelty-P3 amplitudes in the High Quality Team Environment relative to the Low Quality Team Environment, suggesting that cognitive workload was lower in the former. Thus, our metric detected predictable changes in cognitive workload as a function of social environment.

2.4 Summary of Experiments

Collectively, these three studies support our novel metric's ability to assess cognitive workload. Specifically, novelty-P3 amplitude was demonstrated to be sensitive to multiple factors known to influence cognitive workload: changes in task difficulty while holding skill level constant, changes in skill level while holding task difficulty constant, and changes in environmental factors in which both task difficulty and skill level were held constant. Further, in the case of the first two experiments, the metric behaved in a dose-dependent, predictable fashion. Specifically, the metric revealed graded increases in cognitive workload concomitant with incremental increases in task difficulty, and progressive decreases in cognitive workload as a function of skill learning. These results underscore the fidelity and sensitivity of the measure as well as its utility in application.

3 Future Directions

As the employment of this cognitive workload metric progresses, we recommend several future directions regarding research in this area. First, the utility and integrity of this metric need to be rigorously investigated in a myriad of ecologically valid contexts. For example, a study similar to Experiment 1 in the current paper could be conducted in a 'real-world' environment, such as having individuals drive cars during high-density versus low-density traffic. Similarly, the results of Experiment 2 need to be demonstrated to generalize to a diverse set of tasks as a function of learning and

skill level. Secondly, although this metric has been shown to be sensitive to alterations in cognitive workload, a behavioral consequence associated with this index has not been demonstrated. In other words, what is the predictive ability of this metric with regard to performance? For example, one could determine that if the metric suggests an individual is under a high workload, does this correspond to a reduced ability to respond to additional challenge, such as an unexpected, 'surprise,' event. Thirdly, the metric currently requires that novelty-P3 amplitude be determined by computing its average response to multiple stimuli, thus limiting the ability to assess cognitive workload in near 'real time.' Therefore, different signal processing methods (e.g., wavelet analyses) need to be applied in order to compute the novelty-P3 after each stimulus presentation, thereby increasing the temporal resolution of this metric.

4 Conclusion

In this paper we described the utility of a metric that could reliably assess cognitive workload. We then outlined a theoretical rationale for how to assess this and conceived a corresponding novel metric. Experimental evidence was provided that suggested this metric is successful in assessing predictable changes in cognitive workload as a function of task difficulty, learning, and environment. We concluded with recommendations for future research.

References

1. Kahneman, D.: Attention and Effort. Prentice-Hall, Englewood Cliffs (1973)
2. Wickens, C., Kramer, A., Vanasse, L., Donchin, E.: Performance of concurrent tasks: a psychophysiological analysis of the reciprocity of information-processing resources. Science 221, 1080–1082 (1983)
3. Friedman, D., Cycowicz, Y.M., Gatea, H.: The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neurosci. Biobehav. Rev. 25, 355–373 (2001)
4. Isreal, J., Chesney, G., Wickens, C., Donchin, E.: P300 and tracking difficulty: evidence for multiple resources in dual-task performance. Psychophysiology 17, 259–273 (1980)
5. Isreal, J., Wickens, C., Chesney, G., Donchin, E.: The event-related potential as an index of display monitoring workload. Hum. Factors 22, 211–224 (1980)
6. Kramer, A.F., Sirevaag, E.J., Braune, R.: A psychophysiological assessment of operator workload during simulated flight missions. Hum. Factors 29, 145–160 (1987)
7. Sirevaag, E.J., Kramer, A.F., Coles, M.G.H., Donchin, E.: Resource reciprocity: an event-related brain potential analysis. Acta Psychol. 70, 77–97 (1989)
8. Kramer, A.F., Wickens, C.D., Donchin, E.: Processing of stimulus properties: evidence for dual task integrality. J. Exp. Psychol. Hum. 11, 393–408 (1985)
9. Papanicolaou, A., Johnstone, J.: Probe evoked potentials: theory, method and applications. Int. J. Neurosci. 24, 107–131 (1984)
10. Miller, M.W., Rietschel, J.C., McDonald, C.G., Hatfield, B.D.: A novel approach to the physiological measurement of mental workload. Int. J. Psychophysiology 80, 75–78 (2011)

11. Rietschel, J.C., Goodman, R.N., McDonald, C.G., Miller, M.W., Jones-Lush, L., Wittenberg, G.F., Hatfield, B.D.: Psychophysiologcial investigation of attentional processes during motor skill learning. In: 42nd SFN Meeting, New Orleans, LA, USA, October 13-17 (2012)
12. McGill, R.A.: Motor Learning and Control: Concepts and Applications. McGraw-Hill, New York (2011)
13. Miller, M.W., Groman, L.J., Rietschel, J.C., McDonald, C.G., Iso-Ahola, S.E., Hatfield, B.D.: The effects of team environment on attentional resource allocation and cognitive workload. Sport, Exerc., and Perform. Psychol. (2013), doi:10.1037/a0030586
14. Miller, M.W., Presacco, A., Groman, L.J., Bur, S., Rietschel, J.C., Gentili, R.J., McDonald, C.G., Iso-Ahola, S.E., Hatfield, B.D.: The effects of team environment on cerebral cortical processes and attentional reserve. Sport, Exerc., and Perform. Psychol. (in press)

Controlling Attention in the Face of Threat: A Method for Quantifying Endogenous Attentional Control

Bartlett A.H. Russell[1,2,3] and Bradley D. Hatfield[1,2]

[1] Neuroscience and Cognitive Science, University of Maryland, College Park, MD
[2] Department of Kinesiology, University of Maryland, College Park, MD
[3] Center for Advanced Study of Language, University of Maryland, College Park, MD
bartwork@gmail.com, bhatfiel@umd.edu

Abstract. It is well established that anxiety causes attentional narrowing and increases distractibility, yet metrics are lacking for measuring these phenomena during performance. Attention Control Theory (ACT) postulates that anxiety consumes limited executive resources that are necessary for maintaining goal-oriented, "top-down" attentional control and for suppressing stimulus-driven, "bottom-up" distraction. While previous work has quantified the effect of anxious states and traits on bottom-up distraction, it is far more difficult to measure endogenous top-down attention. Here we briefly review theories and previous findings regarding anxiety's affect on attention control and discuss an ongoing study examining sustained attention under neutral and anxiogenic conditions. The study employs a combination of established Electroencephalographic (EEG) methods that together may offer a way to measure top-down sustained attention. If successful, the method could help build a more complete theoretical picture of attention control, and provide a way for HCI platforms to monitor user states in changing contexts.

Keywords: Attention control, Steady State Visual Evoked Potentials, anxiety.

1 Introduction

For human computer interfaces (HCI) to be successful in common real-world uses, they must be robust and reliable despite changing contexts that affect the user physiologically, cognitively or emotionally. This is a particular challenge for HCI that interact with human attention like those that optimize information flow to a user, analyst or operator. Attentional focus is a limited resource and there is often competition between "bottom-up" stimulus-driven attention capture and "top-down" goal-oriented processing. The balance between these two mechanisms enables adaptive filtering that by default directs attention towards potential environmental threats (stimulus-driven), but also permits active inhibition of the stimulus-driven system for goal oriented processing even in the face of distraction. Under normal circumstances this top-down attentional control can override and "tune out" task irrelevant stimuli, but under stress, attentional control is often compromised. Here we discuss the theories and evidence regarding to anxiety's affect on attention control relevant for HCI

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 591–598, 2013.

applications. We outline an ongoing study designed to simultaneously measure bottom-up attentional processing of task irrelevant and task relevant stimuli with Event Related Potentials (ERPs) and continuous top-down attentional control using Steady State Visual Evoked Potentials (SSVEPs) to competing flicker frequencies. We then discuss a few future opportunities and considerations for using such methods in HCI applications.

2 Anxiety and Attention Control

Attention is inherently limited and often characterized as a "spotlight" for highlighting features and cues of interest at the exclusion of others. From a cognitive science perspective, attentional control is a central executive function [1] mediated by bilateral dorsolateral pre-frontal cortex (dlPFC), inferior frontal gyrus, and anterior cingulate cortex (ACC) [2-3]. Areas specific to inhibitory processes thought necessary for suppressing distraction include right ventrolateral prefrontal cortex (vlPFC) and bilateral temporal parietal structures [3]. These frontal regions and their functions are particularly sensitive to affective stimuli, anxiety and stress hormones making them targets for theories and research aimed at understanding top-down attention control.

2.1 Theoretical Models of Arousal and Performance

Anxious arousal is a common feature of many performance environments and is known to undermine the efficiency, speed and/or quality of selective processing of task-relevant information. Theoretical models have evolved over the years to help explain some of these behavioral phenomena. Easterbrook's Cue Utilization Theory [4] was among the earliest of these and attempts to explain why performance may be improved with small amounts of arousal but eventually degrades as arousal reaches maladaptive levels. In short, arousal causes attentional narrowing, initially facilitating performance by excluding task-irrelevant stimuli, in favor of task-relevant information. As arousal increases however, attentional bandwidth narrows to the point that some task-relevant stimuli are also excluded from processing. While useful, Cue Utilization Theory does not explain why performance outcomes are often resistant to anxious arousal, nor why anxious arousal is associated with increased distractibility. Processing Efficiency Theory (PET) offered an explanation to the former question by postulating that arousal and stress affect performance efficiency, if not always performance effectiveness [5]. In stressful contexts individuals can recruit greater neural resources to overcome deficits otherwise associated with attentional narrowing, and maintain the quality of performance at the expense of increased effort. Attention Control Theory (ACT) [6] took PET a step farther to incorporate neurocognitive elements and account for changes in distractibility. Anxious arousal consumes executive resources, eroding inhibitory control allowing salient stimuli – whether relevant or irrelevant – to consume attentional resources [6]. Collectively, these models suggest that under anxiogenic conditions information processing is both limited and inefficiently allocated.

2.2 Evidence Anxiety Erodes Inhibitory Attention Control

Attention Control Theory's (ACT) predictions that anxiety undermines the top-down attentional control and inhibition of task-irrelevant stimuli are supported by experimental findings. While the literature is too vast to review in its entirety here, a few overarching trends and key findings are discussed.

Neurochemical Mechanisms. Attentional control mechanisms have been linked to the function of noradrenaline and dopamine receptors in the prefrontal cortex. Both dopamine (DA) and noradrenaline/norepinephrine (NE) are necessary for enhancing selective attention in the frontal cortex by suppressing neural firing to non-preferred (distracting) stimuli. Stress increases the presence of both chemicals, and too much of either will over-suppress firing, diminishing responses to all stimuli in a non-discriminative manner (for an excellent review see Arnsten, 2009)[2]. This absence of "top-down" prefrontal selectivity putatively increases reactivity to stimulus-driven 'bottom-up" processing and provides a mechanistic basis for understanding the tension between the two systems. Possibly reflecting over-suppression of frontal regions with high DA and NE, fMRI evidence links anxiety-impaired inhibition to decreased activity in attention-related prefrontal brain regions [7]. On a network level, increases in dopamine (induced via a DA reuptake inhibitor) in a resting (non-anxious) state has been linked to the coupling of the frontoparietal control network (FPCN) with the default mode network (DMN) supporting internally-guided attention, and decouples the FPCN from the Dorsal Attention Network (DAN) which otherwise facilitates external cognitive processes [8, 9].

Induced Anxious States. Increasing state anxiety in experimental settings – with threat of shock, psychosocial pressure, or other methods of inducing stress – increases increases sensitivity to bottom-up processing, potentially increasing opportunities for distraction. For example, discrimination between relevant and irrelevant cues was impaired in a driving simulation during competition stress, and indicative of distraction, participants fixated visually more often on peripheral cues [10]. Anti-saccade tasks, which are considered pure reflections of attention control, also show that the speed and efficiency of directing attention *away* from a stimulus are impaired under anxiogeneic conditions [11]. Acute increases in state anxiety increased neural responses to unattended threat stimuli [12] and threat of shock also increased the magnitude of neural response to deviant neutral stimuli indicative of hyper-vigilance [13]. Anxiety manipulations also impair performance on inhibitory tasks: under higher anxiety conditions, participants exhibit slower reaction times in response-conflict tasks [14] and in a dot-probe paradigm [15].

Affective Stimuli. If anxiety increases sensitivity to bottom-up stimulus processing, it is not surprising that threatening or negatively-valenced stimuli garner preferential processing over neutral and positively-valenced stimuli. It is well established that affective stimuli, and particularly negative stimuli, activate the amygdala [16][12] and induce physiological responses, including increased skin conductance and startle responses, even when presented very briefly [17, 18]. Emotional stimuli trigger automatic attention capture mechanisms [19, 20] and show greater hemodynamic

responses in threat-related processing brain regions [21]. Threatening stimuli are also processed faster than neutral or positive stimuli [22][18]. Behavioral results show attentional biases for emotional stimuli in vigilance tasks [23] and across modalities when in competition with simultaneously presented non-arousing stimuli [24].

Trait Anxiety. Other indicators that anxiety impairs inhibitory control are the patterns exhibited by those with high Trait Anxiety. Higher trait anxiety is correlated with increased physiological indices of stress in response to affective stimuli [17] and altered activation patterns in the amygdala [25]. Likewise, increases in trait anxiety measures correlate with reduced attention control and prefrontal activation (dlPFC) and as measured by fMRI [26]. Event Related Potential, (ERP) findings indicate trait anxious individuals exhibit increased attention and neural reactivity to threating stimuli [15], increased responses to deviant stimuli [13], and may also have trouble disengaging with negative stimuli [27], collectively indicating a bias towards threat processing (a potential distraction) and impaired suppression of task-irrelevant stimuli.

Subtypes of anxiety predisposition may also have distinct relationships to neural processes. Those with high social anxiety exhibit increased connectivity between the amygdala and visual processing regions indicating a hypervigilant resting state [28] and a possible bias towards bottom-up, sensory driven pathways. Likewise, while the early N2 amplitude (related to sensory processing) correlated positively with Trait Anxiety scores in an inhibitory task, the amplitude of the subsequent P3 correlated with scores on the Anxiety Sensitivity Index (ASI) [29].

The relevance of these effects of anxiety on attentional systems for human computer interfaces is twofold: 1) optimal information flow will depend partly on the anxious state of the user, which will differ between and among individuals as demands and contexts change; and 2) HCI systems will require means to monitor the user's state if they are to titrate information flow accordingly in response to these changes.

3 An Ongoing Study of Anxiety's Influence on Attention Control

Understanding how anxiety and stress affect attentional control requires simultaneous measurement of bottom–up and top-down attention processes. Of these methods, EEG provides the most practical (non-invasive, inexpensive) option and has the temporal resolution necessary for HCI applications. Indeed, ERPs linked to exogenous stimuli can probe such variables as residual processing capacity [31, 32] and thus provide snapshots of executive function from a bottom-up perspective. Measuring top-down attention control, a dynamic and unpredictable process, presents grater challenges for online assessment.

Steady State Visually Evoked Potentials (SSVEPs) may allow for the persistent, online measurement of endogenous, top-down attention control. SSVEPs are neural oscillations that are induced by flickering stimuli such as a colored shape, checkerboard or Gabor gradient (see Vialatte, et al, 2012 for a current review) [33].[1] If a

[1] These frequency-driven oscillations can also be induced in auditory (Steady State Auditory Evoked Potentials) and somatosensory systems.

flicker is presented at a specific frequency (say, 12Hz), it will drive that same frequency in visual processing regions of the cerebral cortex. The resulting SSVEP is relatively resistant to noise [33] and easily measured with EEG. Most important for studying anxiety's effect on attention control, SSVEPs are attention-sensitive; at certain frequencies in the alpha band (8-12 Hz) SSVEP amplitude is greatest when attention is devoted towards the driving flicker, and is suppressed when the flicker is unattended or actively ignored [34]. These attentional modulations are sensitive to covert shifts in attention and do not require visual fixation on the flicker [35] meaning top-down attention modulation can be isolated from sensory-dependent bottom-up systems. Because the visual flicker generates a persistent frequency tag in the neural tissue, deviations of this tag may allow us to quantify and detect unpredictable endogenous changes in attention control (such as attentional lapses or "zoning out") in a way that snapshot methods cannot.

Morgan, Hansen & Hillyard, (1996) [35] demonstrated that more than one SSVEP frequency can be stimulated and recorded at the same time and that the relative amplitude to each frequency reflects covert attention allocation. In other words, if attention is shifted from frequency A to frequency B, the amplitude of SSVEPA will decrease and the amplitude of SSVEPB will increase relative to baseline. This sort of experimental setup allows for the simultaneous, persistent measurement of to-be-ignored and to-be-attended stimuli; greater amplification of the attended SSVEP and suppression of the ignored SSVEP are indicators of more selective sensory filtering by top-down mechanisms [36]. Embedding targets and distractors in the competing stimulus flickers generate attention-dependent ERPs providing simultaneous snapshots of bottom-up attention capture by task-relevant and task-irrelevant stimuli.

We are using the same experimental approach to assess how anxiety (unpredictable threat of shock) affects covert attention control in a target detection task. If anticipatory anxiety erodes one's ability to ignore task-irrelevant stimuli, individuals should exhibit: 1) less SSVEP suppression of an unattended flicker; and 2) increased attentional capture by distractor stimuli presented within the unattended flicker compared to neutral conditions.

SSVEPs have been long used and developed for use in brain-computer-interfaces (BCI) as a way to encode user commands without relatively slow motor and mechanical intermediaries [33], and thus their utility in HCI is already established. If the above described methodology proves sensitive to such manipulation, BCI and HCI systems could also use SSVEPs for the persistent, online assessment of a user's attention bandwidth and vulnerability to distraction - whether anxious, fatigued, or otherwise compromised – to optimize and titrate information flow. For civilian and military applications such a method could provide a means for: assessing qualitative differences among individuals who perform either very well or very poorly in stressful situations during training and selection; quantifying the effectiveness of training for reducing stress susceptibility; designing and testing platforms that may offset the deleterious effects of anxiety; and monitoring the operator's ability to control his or her attention during performance.

4 Final Thoughts and Future Opportunities

While the results of the ongoing study are pending, other features of SSVEPs present a number of potential opportunities and a few challenges for HCI applications. For example, different frequencies drive SSVEPs in distinct populations of neural tissue, and may allow for network- or population-specific frequency tagging. While alpha band frequencies (8-12Hz) are attention sensitive, upper and lower alpha frequencies show distinct responses, and other frequencies seem to be unaffected by attention [37] An HCI system could leverage this specificity to monitor and parse distinct neural processes and networks with functionally distinct frequencies. Measures of phase – while partially reflected in measures of amplitude – may also be of utility. Despite this potential for network specificity, the presentation of a series of flickers can be visually fatiguing. As a result computer interfaces looking to include such tools will have to be selective in choosing the frequency and stimuli that will provide the most robust signal, while minimizing perceptual demand. In sum, and despite such difficulties, the example of using a combined ERP and SSVEP paradigm to investigate the effects of anxious states on attention control mechanisms illustrate how this method could provide a more complete picture of attentional control mechanisms for theoretical and functional understandings of these systems as well as improved tools for optimizing human computer interaction.

References

1. Baddeley, A.: Exploring the central executive. The Quarterly Journal of Experimental Psychology: Section A 49(1), 5–28 (1996)
2. Arnsten, A.F.T.: Stress signaling pathways that impair prefrontal cortex structure and function. Nature Reviews Neuroscience 10(6), 410–422 (2009)
3. Hedden, T., Gabrieli, J.D.E.: Shared and selective neural correlates of inhibition, facilitation, and shifting processes during executive control. NeuroImage 51(1), 421–431 (2010)
4. Easterbrook, J.A.: The effect of emotion on cue utilization and the organization of behavior. Psychological Review 66(3), 183–201 (1959)
5. Eysenck, M.W., Calvo, M.G.: Anxiety and Performance: The Processing Efficiency Theory. Cognition & Emotion 6(6), 409–434 (1992)
6. Eysenck, M.W., Derakshan, N., Santos, R., Calvo, M.G.: Anxiety and cognitive performance: Attentional control theory. Emotion 7(2), 336–353 (2007)
7. Bishop, S.J.: Neurocognitive mechanisms of anxiety: an integrative account. Trends in Cognitive Sciences 11(7), 307–316 (2007)
8. Spreng, R.N., Stevens, W.D., Chamberlain, J.P., Gilmore, A.W., Schacter, D.L.: Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. NeuroImage 53(1), 303–317 (2010)
9. Dang, L.C., O'Neil, J.P., Jagust, W.J.: Dopamine supports coupling of attention-related networks. Journal of Neuroscience 32(28), 9582–9587 (2012)
10. Janelle, C.M.: Anxiety, arousal and visual attention: a mechanistic account of performance variability. Journal of Sports Sciences 20(3), 237–251 (2002)
11. Ansari, T.L., Derakshan, N.: The neural correlates of impaired inhibitory control in anxiety. Neuropsychologia 49(5), 1146–1153 (2011)

12. Bishop, S.J., Duncan, J., Lawrence, A.D.: State anxiety modulation of the amygdala response to unattended threat-related stimuli. Journal of Neuroscience 24(46), 10364–10368 (2004)
13. Cornwell, B.R., Baas, J.M.P., Johnson, L., Holroyd, T., Carver, F.W., Lissek, S., Grillon, C.: Neural responses to auditory stimulus deviance under threat of electric shock revealed by spatially-filtered magnetoencephalography. NeuroImage 37(1), 282–289 (2007)
14. Choi, J.M., Padmala, S., Pessoa, L.: Impact of state anxiety on the interaction between threat monitoring and cognition. NeuroImage 59(2), 1912–1923 (2012)
15. Eldar, S., Yankelevitch, R., Lamy, D., Bar-Haim, Y.: Enhanced neural reactivity and selective attention to threat in anxiety. Biological Psychology 85(2), 252–257 (2010)
16. Hariri, A.R., Mattay, V.S., Tessitore, A., Fera, F., Weinberger, D.R.: Neocortical modulation of the amygdala response to fearful stimuli. Biological Psychiatry 53(6), 494–501 (2003)
17. Smith, J.C., Löw, A., Bradley, M.M., Lang, P.J.: Rapid Picture Presentation and Affective Engagement. Emotion 6(2), 208–214 (2006)
18. Harald, T., Schupp, H.T., Junghöfer, M., Weike, A.I., Hamm, A.O.: The Selective Processing of Briefly Presented Affective Pictures: an ERP Analysis. Psychophysiology 41(3), 441–449 (2004)
19. Carreti, L., Hinojosa, J.A., Martín-Loeches, M., Mercado, F., Tapia, M.: Automatic attention to emotional stimuli: Neural correlates. Human Brain Mapping 22(4), 290–299 (2004)
20. Thierry, G., Roberts, M.V.: Event-related potential study of attention capture by affective sounds. Neuroreport 18, 245–248 (2007)
21. Bishop, S.J.: Neural Mechanisms Underlying Selective Attention to Threat. Annals of the New York Academy of Sciences 1129(1), 141–152 (2008a)
22. Schupp, H.T., Öhman, A., Junghöfer, M., Weike, A.I., Stockburger, J., Hamm, A.O.: The Facilitated Processing of Threatening Faces: An ERP Analysis. Emotion 4(2), 189–200 (2004)
23. Carrettie, L.: Valence-related vigilance biases in anxiety studied through event-related potentials. Journal of Affective Disorders 78(2), 119–130 (2004)
24. Keil, A., Bradley, M.M., Junghöfer, M., Russmann, T., Lowenthal, W., Lang, P.J.: Cross-modal attention capture by affective stimuli: evidence from event-related potentials. Cognitive, Affective, & Behavioral Neuroscience 7(1), 18–24 (2007)
25. Davidson, R.J.: Anxiety and affective style: role of prefrontal cortex and amygdala. Bps 51(1), 68–80 (2002)
26. Bishop, S.J.: Trait anxiety and impoverished prefrontal control of attention. Nature Neuroscience 12(1), 92–98 (2008b)
27. Koster, E., Crombez, G., Verschuere, B., Vandamme, S., Wiersema, J.: Components of attentional bias to threat in high trait anxiety: Facilitated engagement, impaired disengagement, and attentional avoidance. Behaviour Research and Therapy 44(12), 1757–1771 (2006)
28. Liao, W., Qiu, C., Gentili, C., Walter, M., Pan, Z., Ding, J., Zhang, W., Gong, Q., Chen, H.: Altered Effective Connectivity Network of the Amygdala in Social Anxiety Disorder: A Resting-State fMRI Study. PLoS ONE 5(12), e15238 (2010)
29. Sehlmeyer, C., Konrad, C., Zwitserlood, P., Arolt, V., Falkenstein, M., Beste, C.: ERP indices for response inhibition are related to anxiety-related personality traits. Neuropsychologia 48(9), 2488–2495 (2010)
30. Huang, Y., Bai, L., Ai, H., Li, W., Yu, C., Liu, J., Luo, Y.J.: Influence of trait-anxiety on inhibition function: Evidence from ERPs study. Neuroscience Letters 456(1), 1–5 (2009)

31. Miller, M.W., Rietschel, J.C., McDonald, C.G., Hatfield, B.D.: A novel approach to the physiological measurement of mental workload. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology 80(1), 75–78 (2011)
32. Rietschel, J.C., Goodman, R.N., McDonald, C.G., Miller, M.W., Jones-Lush, L.: Psychophysiologcial investigation of attentional processes during motor skill learning. In: 42nd SFN Meeting, New Orleans (2012)
33. Vialatte, F.B., Maurice, M., Dauwels, J., Cichocki, A.: Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives. Progress in Neurobiology 90(4), 418–438 (2010)
34. Silberstein, R.B., Schier, M.A., Pipingas, A., Ciorciari, J., Wood, S.R., Simpson, D.G.: Steady-state visually evoked potential topography associated with a visual vigilance task. Brain Topography 3(2), 337–347 (1990)
35. Morgan, S.T., Hansen, J.C., Hillyard, S.A.: Selective Attention to Stimulus Location Modulates the Steady-State Visual Evoked Potential. Proceedings of the National Academy of Sciences of the United States of America 93(10), 4770–4774 (1996)
36. Mishra, J., Zinni, M., Bavelier, D., Hillyard, S.A.: Neural Basis of Superior Performance of Action Videogame Players in an Attention-Demanding Task. Journal of Neuroscience 31(3), 992–998 (2011)
37. Ding, J., Sperling, G., Srinivasan, R.: Attentional Modulation of SSVEP Power Depends on the Network Tagged by the Flicker Frequency. Cerebral Cortex 16(7), 1016–1029 (2005)

Developing Visualization Techniques for Improved Information Comprehension and Reduced Cognitive Workload

Scott Scheff[1], Tristan Plank[1], John Wilson[1], and Angelia Sebok[2]

[1] HF Designworks, Inc.
[2] Alion Science and Technology
scottscheff@hfdesignworks.com

Abstract. In today's data rich environments, enormous quantities of digital information can now be collected and made available to end-users in a wide variety of domains. With so much information now readily accessible, effective display methods that integrate and make sense of the data are needed; otherwise end-users may quickly become overwhelmed. HF Designworks, Inc. and Alion Science & Technology have developed tools that leverage large quantities of information to provide useful visualizations to the warfighter. This paper describes the approach and results of two related projects, iWarrior and My Heat Maps, where we provide end-users with deep data comprehension without imposing cognitive overload.

Keywords: Applications of Augmented Cognition.

1 Introduction

Military, commercial, and medical sectors now typically yield enormous quantities of data for personnel to interact with and interpret. For example, typical modern military areas of operation continuously collect data from advanced sensors and other intelligence-gathering tools. These large amounts of information can offer valuable insights into a variety of battlefield contexts. However, for this information to be usable, novel and effective techniques are required in order to sort, filter, and display data to end-users without overwhelming them. Without effective display techniques, high volumes of data can become unusable and potentially hinder military operations by pulling time and manpower from needed areas. A prime example of a fielded system that yields high quantities of data is "blue force tracker" (BFT) situational awareness data from the battlefield. BFT data provides the current and recent past location of military ground-based assets, and is vital for logistics, mission planning, and identifying gaps in strategy and area coverage. Yet much of this valuable information is not leveraged to its full potential because of the sheer quantity and format of data. With Blue Force GPS data, a single vehicle traveling for a few hours can generate thousands of "GPS footprints," which must then be plotted and mapped before they become useful. When tracks of multiple vehicles are gathered, the data quantity multiplies, as does

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 599–607, 2013.

the difficulty in interacting with the data. However, the potential usefulness of data also multiplies as patterns begin to emerge in the frequency, density, and gaps in coverage of tracks over a geographical region. To address this potential for data overload, applications are being developed to provide enhanced visualizations that expose patterns in travel behavior. Such patterns are important to warfighters because frequently traveled routes may be more likely to be observed by the enemy, and are of greater interest to the enemy, possibly resulting in attacks and ambushes along that particular route. Additionally, regions with little traffic may indicate seldom-patrolled locations where the enemy could convene or find a safe haven.

2 Approach

2.1 iWarrior

With iWarrior, HF Designworks, Inc. and Alion Science & Technology (the team) seek to provide end-users with deeper and more complete data comprehension while avoiding the cognitive overload that can quickly arise when interacting with large quantities of data. Funded as part of a Defense Advanced Research Projects Agency (DARPA) Small Business Innovative Research (SBIR) effort, iWarrior is a web-based tool for supporting battle space awareness and planning. iWarrior collects historical GPS coordinates (from military location data-tracking programs such as TiGR, FBCB2, or other sources with a minimum latitude/longitude and a Date Time Group) and uses this data to produce map-based visualizations of Soldier and vehicle tracks. iWarrior also provides route analysis capabilities with tools such as “heat map” overlays. Heat maps use coloring scales to display relative traffic in geographic regions, visually indicating heavily traveled, “hot” areas (e.g., red), moderately traveled “warm” areas (e.g., orange or yellow) or seldom traveled “cool” areas (e.g., white or light blue), Fig.1.

Fig. 1. View of iWarrior Density Visualization

iWarrior was developed through a multi-year effort that involved an iterative approach of working with subject matter experts, identifying information requirements, working with technical experts to specify what can be implemented, and performing rapid prototyping to display and test concepts. The approach for iWarrior development has been focused on user requirements; use case-based development, and frequent user testing. The iWarrior tool emphasizes practical solutions for addressing user concerns and presenting accurate representations of the data in an effort to not mislead users.

2.2 My Heat Maps

The heat mapping capabilities and algorithms explored in iWarrior have also been developed for Android-based handheld devices through the DARPA Transformative Applications (TranApps) program. HF Designworks has developed a plug-in for the TransApps 'Maps' application that converts the handheld device's recorded GPS tracks to heat maps that show the traffic density in an area, Fig.2. This capability allows Soldiers to view their travel behavior in the form of visual patterns, allowing them to readily note points of vulnerability (either frequently-traveled areas, or locations on their patrol that have been neglected and may need additional presence). These capabilities are all available directly from the handheld device without needing to rely on data connectivity. In other words, all processing is performed on the device itself. These heat-mapping capabilities also have potential for use in a variety of other applications. For instance, we have recently applied these mapping capabilities to a civilian Unmanned Aircraft System (UAS), marking where that aircraft has flown and with what frequency the area was covered. Future applications could include heat mapping not only the paths themselves of UAS deployed in environments such as combat zones, search and rescue, and land surveying, but also the area coverage captured by sensor payloads to reveal locations that have been missed in flyovers. This would allow operators to maximize area coverage, minimize fuel consumption, and avoid retracing areas unnecessarily.

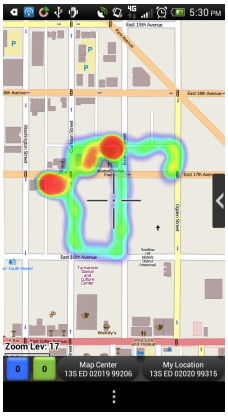


Fig. 2. View of a heat map rendered on the device

3 Requirements

A past project completed by the team demonstrated that many of the military casualties that occur during a tour of duty are due to complacency that can lead to selecting and traveling along the same route repeatedly or conducting patrols in a predictable manner. As a result, enemy forces detect patterns in the patrols or routes used, and use this knowledge to stage attacks or select locations for Improvised Explosive Device (IED) placement. Based on this finding, a tool was needed that could easily and effectively display tracks that show where vehicles have traveled for specific periods of time, presented on a map. Track-based visualizations are the core features of both iWarrior and My Heat Maps, and almost all other features within these tools build from this.

Starting with a focus on route traffic, various visualization features were developed through SME interviews and iterative designing. In addition, the development of heat mapping visualizations (which use coloring scales to display relative traffic in geographic regions, visually indicating heavily traveled, "hot" areas), Soldiers also indicated a need for identifying areas where vehicles or personnel have remained stationary for longer periods of time. This requirement resulted in the development of halt visualizations to represent where vehicles or personnel have remained stationary (based on GPS tracks recording periods of no movement) for a user specified amount of time. The GPS data is also used to provide statistics and metric visualizations to users so they can better determine how fast units move through selected areas as well as the frequency of movement based on time (year/month/day and even time of day).

Continual user feedback and SME-derived requirements resulted in a number of other features; iWarrior now also provides push pins for identifying significant events associated with a specific location or a selected region on the map. Users place pins and provide associated data (e.g., text or image files). Pins offer visual indications of vital information, allowing warfighters to recognize possible relationships between GPS traffic and push pin data.

User requirements have also guided a recent update to the My Heat Maps application which enables users to save their tracks as heat map image files which can then be uploaded to a computer and viewed in Google Earth, offering an additional method for tracking route and patrol traffic as well as allowing users to combine tracks from multiple handsets for viewing in Google Earth. Note that the popularity of viewing image files in Google Earth by Soldiers has also allowed us to implement a similar image conversion and download feature in iWarrior.

4 Prototype

4.1 iWarrior

iWarrior's features have been developed to enhance warfighter comprehension of the battle space and augment decision-making through valuable visualizations and information management tools, as was presented in Fig.1. The use of iWarrior begins with

selecting a segment of time (by choosing a start and end time) or a geographic area on the map. The tool then retrieves the GPS data for this period of time/location and plots the tracks on the map interface, allowing the data to be 'played back' at various speeds. The interface includes a variety of filters (developed based on user needs and SME feedback). These filters allow users to quickly view and hide information overlays, such as turning off the Density visualization to view only GPS tracks, turning on and off track bounds to help find areas of heavy traffic when zoomed out, and turning push pins on and off. The interface is designed in "layers," which allow the user to select and display different types of data. This lets the user compare items of interest (e.g., density and push pins) without cluttering the display unnecessarily. The layers provide users the ability to switch between (or display simultaneously) the map itself, the GPS track layer, Density layer, push pins layer, or track bounds layer. Map tools for measuring distance and area are also included. A background fader has also been developed which allows users to lighten or darken the map background to improve the clarity and contrast of tracks and densities, which also are supported in multiple color schemes. Due to its popularity, this fader feature has been carried over to other map based programs as well. Features such as density, clusters, push pins, the background fader widget, and a statistics tool (density by time of day in this example) are shown in Fig.3.

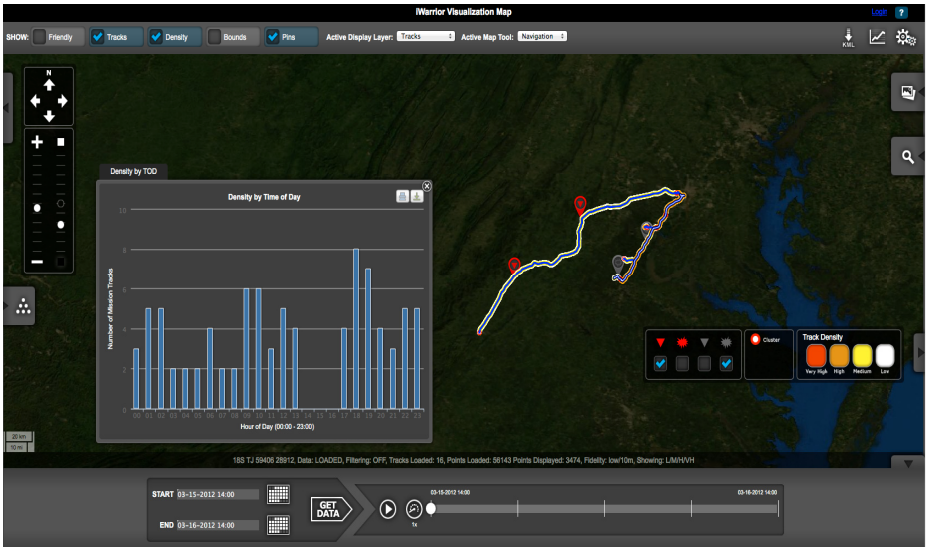


Fig. 3. iWarrior with Density, Clusters, Push Pins, Background Fader Widget, and Time of Day Density Features All Active

4.2 My Heat Maps

My Heat Maps is a plugin to the DARPA Transformative Applications (TransApps) Maps application. As with iWarrior, heat maps are based on GPS tracks; therefore, users create 'sessions' of GPS data using the device's built-in GPS. These sessions

are populated in the My Heat Maps plugin, and users then select sessions they would like to generate into heat maps. My Heat Maps can also convert the heat maps and save them on the hand held's SD card, allowing them to be uploaded to a computer and viewed in Google Earth (Fig.4) which allows for creating presentations for pre and post mission as well as collecting and displaying heat maps for multiple users.

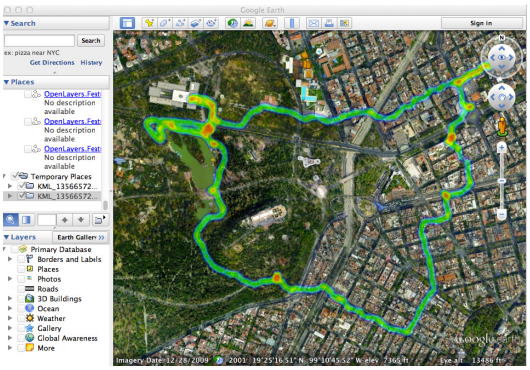


Fig. 4. A heat map image, converted, exported and viewed in Google Earth

5 Testing of Concepts and Results

The development of both iWarrior and My Heat Maps is an iterative process including periods of test-fix-test. As such, we have tested both the iWarrior and handheld applications, and identified results that are continually used to improve these applications. Results are described below.

5.1 iWarrior

iWarrior has been tested internally with our SMEs (all with recent combat experience). Additionally, iWarrior features such as density heat maps and statistics have been integrated into larger Department of Defense (DOD) information systems and tested both by SMEs as well as evaluated through limited field-testing. Utilizing this type of testing as part of our iterative design process allows us to continually improve our visualization tools. Working with end users also allows us to receive feedback on desired features, which we can then incorporate into our products.

At this time we do not have formal test results from our iWarrior tool. Initially, based on Phase I interviews and focus groups we recognized a need for such a tool. iWarrior was then developed with Soldiers as SMEs assisting us throughout the design process with the goal to eventually test iWarrior through simulated, SME-based missions .

5.2 My Heat Maps

To explore the in-field applicability of the My Heat Maps application plugin, a four-man team comprised of our SMEs performed a field test to gather GPS data of possible real-world scenarios. All four members of the team had prior military experience including several deployments to Afghanistan and Iraq as infantrymen. The main goal of the exercise was to see how the on-the-ground movements translated into heat maps and then use the heat map visualizations to assist Soldiers in conducting future operations. To accomplish this goal, the four-man team conducted a variety of basic military formations and common mission scenarios while recording their GPS data.

The results of the heatmap evaluation showed us that we were on the right track in terms of visually indicating to users where they had been, what type of movement was made, and with what frequency and speed they had moved through select areas. Results and later discussions also indicated to us that we needed to heat by mission rather than track if the heatmaps for multiple personnel were being viewed at the same time. Displaying heatmaps for multiple personnel can be performed via uploading to a computer and displaying in Google Earth, or uploading data to iWarrior (a single hand held device using My Heat Maps only renders heat maps for a single user). Additionally, because some units may have several hand held devices while others might only have a single hand held device, we would need to revise our algorithms so that each mission, regardless of how many tracks were included, received one heat value. The more missions in the same area would result in a higher heat value. This was because we were making the assumption that the enemy was more interested in the fact Soldiers travelled through an area and how often, rather than how many Soldiers travelled through that area (which could be confusing to the system since there is no way to know in advance just how many Soldiers would have hand held devices).

The team found that heat maps were an easy and effective way of visualizing routes taken. The heat maps also helped identify choke points in areas of slow movement and areas covered in clearing operations. The team agreed that the information from heat maps would be very useful for mission planning, after action reviews and debriefs. The team also suggested that heat maps could be improved by providing additional capabilities. Capabilities that could enhance the utility of the program include: the ability to vary the rate of 'heating', the ability to vary the thickness of heat maps for different terrain (e.g. jungle vs. desert environments), introducing improved methods for indications movement speeds, and an option to have heat maps fade overtime (i.e., stale out).

For future work we plan to show these heat maps to SMEs who were not part of the original missions. We will then asking these new SMEs to evaluate the heat maps and provide feedback on what they think is being presented (i.e., movement details), to help ensure our heating algorithms are fully accurate over a variety of environments and mission scenarios.

Fig.5 and Fig.6 display the heat mapping data of four users traveling in a fire team formation on line. Fig.5 shows the fire team as they spread out, moving across the image area from left to right. Fig.6 shows the fire team spreading out and then collapsing back in as they move across a field.

Fig. 5. Fire Team's heat map as they spread out, moving from the left to right side of the image

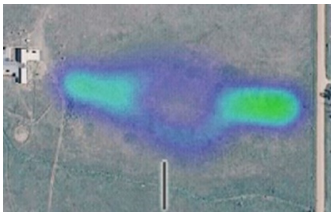


Fig. 6. Fire Team's heat map as they spread out and then collapse back together as they move across a field from the left to right side of the image

Fig.7 shows a heat map for a patrol starting at point A and moving to point B *(The screenshot has an overlay to clarify movement).* The heat map visualization allows Soldiers to easily determine where they have been in order to plan future missions to maximize area coverage and reduce predictable routes.

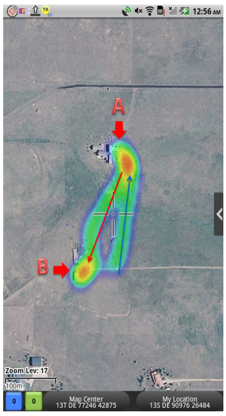


Fig. 7. Heat Map of a Patrol scenario starting at Point A and patrolling to Point B. The arrows indicate the different routes taken to and from Point B.

6 Next Steps

We will continue to work with the warfighter, identifying methods for evaluating data then using that information in an intuitive manner to help the warfighter with their tasks. This can include improving heat mapping (adding the ability to customize how

and when GPS points get heated) or even developing cultural conversion applications so Soldiers can better understand foreign communities they encounter. Further testing will include mission-based scenarios to gather data by one team of SMEs, and reviews/interpretation by a second group of SMEs. When the second SMEs have developed their understanding, we will have them present to the first group. This will allow us to identify potentially misleading data visualizations, and to develop solutions that make sense to all users.

7 Summary

Current and rapidly emerging technologies can provide users an abundance of vital data. The data itself however, is not necessarily useful until it is displayed in a way that end-users can comprehend. With visualization techniques that organize data in a meaningful format for the intended end user, digital data can be used to provide key information to users without contributing to, or causing, information overload. In the cases of iWarrior and My Heat Maps we were able to successfully take historical GPS data and enhance a user's situation awareness and decision making through statistical charts (including showing critical times of day Soldiers are conducting missions) as well as heat maps of what areas are being covered and with what frequency.

Additionally, by following an iterative design cycle where end users are involved throughout, including a final test-fix-test cycle at the end of each build, we can further ensure the products meet the needs of the users and are robust enough to work, and work well, in the field.

References

1. Milanski, J., Scheff, S.: Usability of a Body-Borne Computer for US Army Infantrymen on the Move. Journal for the International Society for Occupational Ergonomics & Safety (2005)
2. Plank, T., Scheff, S., Sebok, A.: Capturing Insights to Reduce Future Warfighter Fatalities. Human Factors and Ergonomics Society (2010)
3. Plank, T., Scheff, S., Sebok, A.: DARPA-Commissioned Study Investigates Soldier Survivability, Finds First 100 Days of Deployment Critical to Soldier Survivability. National Defense Magazine (2010)
4. Scheff, S., Katz, J.: Human Factors Design for the US Army Infantryman. Journal of Human Performance in Extreme Environments (2004)

Development of Fatigue-Associated Measuresment to Determine Fitness for Duty and Monitor Driving Performance

Ying Ying Tan, Sheng Tong Lin, and Frederick Tey

Combat Protection & Performance Program, DSO National Laboratories
27 Medical Drive #09-00, Singapore 117510
tyingyin@dso.org.sg

Abstract. Long distance driving has been a major factor leading to road accidents [1-2]. With the lack of reliable validation on driver fatigue technology systems [3], the aim of this study is to correlate the measurements of two cognitive tests: Psychomotor Vigilance Task Tester-PVT [4] and PenScreen-PS [5] to establish the threshold levels of fatigued driving performance that will form the basis to prevent fatigued drivers from handling vehicles. PVT is recommended to be the first line of defense against putting fatigued drivers on duty. Drowsiness can be detected by SmartEye Anti-Sleep-AS, acting as a monitoring tool. Eye closure analysis on AS's eyelid opening data showed that AS is a feasible system for real-time monitoring of fatigue while driving. The results also suggested a simpler and more economical way of monitoring fatigue using AS system. PS could be used in conjunction with PVT to detect for any malingering intent.

Keywords: Fatigue, Fitness for Duty, Driving Performance.

1 Introduction

Long haul driving is an example of a prolonged operation or task that demands sustained vigilance in which human performance eventually breaks down as a result of mental fatigue. This can cause safety to be compromised. Operator fatigue has been one of the most prevalent reasons behind accidents, even in military settings, leading to the development of Fatigue Management Technologies (FMT) [6]. Generally, the fatigue problem is tackled by these FMTs in two ways.

One of the ways to mitigate driver fatigue is to monitor fatigue real-time and indicate its onset through a warning system. Such monitoring systems have the added advantage of measuring driver's alertness while he drives, without requiring him to perform additional and possibly, distracting tasks. However, current technology is limited to detecting the onset of fatigue instead of predicting it, and hence does not allow for early intervention. Additionally, current behavioral attributes monitored are largely controllable by conscious means. In other words, unmotivated operators can

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 608–617, 2013.

mimic fatigue-like behaviors to trick the fatigue-warning system, so as to be excused from mandatory duties. One such system is the PERCLOS system that measures the percentage of eyelid closure to infer sleeping behavior.

Therefore, there is a need to develop a robust early predictor by monitoring attributes that cannot be voluntarily controlled by the observed driver. Monitoring involuntary attributes like saccadic eye velocity (quick eye movement speed) and pupil reflexes seems to be a better approach and these actions have been found to be highly correlated to fatigue levels [7-12]. Heart rate variability has been found to be a useful covariate of fatigue [13-17] as well as electrodermal activity (EDA), which detects the changes in skin activities [18, 19]. This study aims to validate eye reflexes, heart rate and EDA measures as effective early predictors for unacceptable fatigue levels.

This study would potentially lead to improvements in operation safety of extended operations and sustained demand for vigilance by preventing human errors due to fatigue. Furthermore, the detection concepts developed here have the advantage of not requiring the driver to perform additional tasks which can be a hassle to the driver and potentially detract him from his primary task.

2 Method

Forty healthy Singaporean male participants (aged 20 - 45) licensed to drive a motor vehicle weighing no more than 3000 kg with no bad driving records for the past one year were recruited. All interested and eligible participants attended a recruitment brief at least three days ahead of their trial. During the brief, details on the conduct of the trial, trial safety aspects, and subject reimbursement were presented. Participants willing to take part in the trial signed an informed consent in the presence of a witness (minimum 21 years of age). Each of them was issued an ActiWatch, a wrist-device to log their sleep duration for 3 days before his trial. This study required participants to have minimum 6 hours of sleep every night, for 3 nights, prior to their trials.

Informed consent, indemnities and recruitment work processes was administered to those interested on the same day, less those who are below 21 years of age and require parent's consents.

The fatigue driving trial required participants to perform prolonged monotonous driving (30km/hr, up to a maximum of 4 hours) within a closed-circuit road (refer to Table 1). Three cognitive systems deployed to determine the participants' pre–post fatigue driving differences. A monitoring system tracked the participants during the entire driving duration.

Each participant was required to complete both Trial A and B on separate week-ends at least 6 days apart to prevent fatigue interaction between trials. The sequence of trials was counterbalanced between two groups of participants. Trial A was de-signed to apply cognitive test hourly during the 4 hours driving, while in Trial B, cognitive tests were administered pre and post driving.

Table 1. Work flow for Trial A & B

Time Start	**Activities**	**Equipment (Trial A)**	**Equipment (Trial B)**
0730 hrs	Reporting and Safety Briefing	All equipment ready. ActiWatch data to be downloaded. Health Declaration signing.	
0750 hrs	Equipment Familiarization	PVT, PS – Demo	
0830 hrs	Driving Familiarization	On-the-road Driving 5 Rounds	
0840 hrs	Breakfast	Food	
0915 hrs	Measurement #0 – Trial Baseline	PVT, PS, VAS	
0930 hrs	Driving Game	PS3 Game console	
1130 hrs	Measurement #1 – Before Driving	PVT, PS, VAS	
1205 hrs	Light lunch	Food	
1215 hrs	Baseline Driving	ET(AS),	
1230 hrs	Driving 30km/hr for 4 hours*	Continuous data collection – ET(AS), Hourly stoppage for PVT, PS, VAS	Continuous data collection – ET(AS),
1630 hrs	Measurement #2 – After Driving	PVT, PS, VAS	
1700 hrs	Monotonous Driving Game#	Continuous data collection - ET(AS), PS3 Game console	Continuous data collection - T(AS), PS3 Game console
1900 hrs	Measurement #4 – After Game	PVT, PS, VAS	
1815 hrs	End of Trial-Run	Pack all equipment	

PVT – Psychomotor Vigilance Task Tester

PS – PenScreen

VA – Visual Analogue Scale (Fatigue Survey)

EDA – Electro-Dermal Activity

ET(AS) – Eye tracker (Smart Eye Anti-Sleep)

Note: *Participant will proceed to the next item if he dozes off less than 4 hours into driving.
#Participant will proceed to end the trial if he dozes off less than 2 hours into the monotonous driving game.

3 Results and Discussion

Actiwatch II – Participants, on average, rested 441.22 minutes per night for 6 nights prior to taking part in the driving trials. The Actiwatch II registered 91.2% of these times as sleep times. On the night before the trial, participants rested, on average, 384.11 minutes. No significant correlations between rest/sleep duration and driving duration or cognitive task performance.

3.1 Visual Analogue Scale [VAS]

On average, participants in Trial A drove for 151 minutes while participants in Trial B drove for 149 minutes. The analysis on VAS showed that the participants' perception of fatigue increased significantly after the driving task. This supports the claim that the driving task successfully induced fatigue. The duration of driving in Trial B did not correlated with the increase in VAS score, meaning with a longer period of driving did not corresponded to a proportionate increase in the participants' rating in fatigue and that the two measures were independent of each other.

3.2 Grouping of Participants (Refer to Table 2)

From the results derived from Psychomotor Vigilance Test (PVT) and PenScreen (PS), serving as potential screening tool, the performance of these participants were classified into three groups. One group of participants (n = 11, were labeled as Elites), all the participants drove for more than 220 minutes for both Trial A and B. They were able to successfully complete the full driving task without lane deviation. Another group of participants (n = 14) could only drive less than 90 minutes for both trials were known as the Vulnerable drivers, as they could not complete the full 4-hour driving task and had to be stopped for causing danger to other road users. Seventy-five percent of this group of participants for managed to drive for more than 40 minutes. The longest driving duration in the group was 160 minutes and the shortest driving duration in the group was 18 minutes. The last group of participants (n = 15), better known as the Malingerers, managed to drive for more than 140 minutes for both trials on average before lane deviating. They have the tendency to drive almost 90 minutes longer during the first trial and most of the time participants in this group did not lane deviate in the first trial, but lane deviated in the second trial.

In summary, it was believed that the participants in this group purposefully drove less on their second trial than on their first to try to end the trial earlier, regardless of the different task demands required for each trial. We call this group the Malingerers, as they seemed to have feigned fatigue to get off the driving trial when we suspect they have not reached their maximum fatigue level. 13 out of the 15 Malingerers, who drove for more than 180 minutes, did it only for their first trial but not for the second trial. The longest driving duration for the group was 242 minutes while the shortest driving duration was 16 minutes.

Table 2. Grouping criteria for elite, vulnerable and unmotivated drivers

	Elites	Vulnerable	Unmotivated
First Trial	Can last approximately 4 hours of driving without deviating from lane.	Can only last approximately 1 hour of driving without deviating from lane.	Can last approximately 4 hours of driving without deviating from lane.
Second Trial			Can only last approximately 1 hour of driving without deviating from lane.

Sleep records obtained from the Actiwatch found no significant differences between these 3 groups of participants in terms of rest and sleep duration ($F(1,36) = 0.18$, $p = 0.836$). Thus, all significant differences that are found from analyses that follow cannot be due to the effects of rest and sleep.

3.3 Cognitive Test –Psychomotor Vigilance Test (PVT)

The results from the Vulnerable group revealed that PVT can reliably screen for fatigue individuals who are unfit for road duties. The following tables described the significant differences between and within group comparison for PVT mean reaction time and its standard deviation.

Table 3. PVT mean RT (ms) with lapses

Trial	Time	Elite	Vulnerable	Malinger	Comparison by Group	Comparison by Time
A	Start	249.83 (28.51)	280.66 (55.05)	265.70 (39.44)	***F*(2,37) = 5.17 p <0.05 (Elite vs.**	*F*(1,37) = 27.57 p<0.01
	End	279.20 (60.36)	560.77 (282.72)	450.15(215.12)		
Comparison within groups		*F*(1,10)= 4.29 p = 0.065	***F*(1,13)=15.72 p < 0.01**	***F*(1,14)= 14.32 p < 0.01**	Time*Group Interaction ***F*(2,37)=5.04, p < 0.05**	
B	Start	264.75 (32.51)	264.60 (31.90)	289.58 (68.23)	*F*(2,37) = 0.74 p = 0.483	***F*(1,37) = 5.13 p<0.05**
	End	295.54 (52.58)	505.90 (337.94)	499.01 (704.58)		
Comparison within groups		***F*(1,10)=8.63 p < 0.05**	***F*(1,13)= 7.28 p <0.05**	*F*(1,14)= 1.66 p = 0.218	Time*Group Interaction *F*(2,37)=0.77, p = 0.471	

Note: values in brackets are Standard Deviation of its respective mean.

Table 4. Standard deviation (SD) of RT with lapses

Trial	Time	Elite	Vulnerable	Malinger	Comparison by Group	Comparison by Time
A	Start	61.49 (30.04)	137.37 (186.86)	73.88 (46.84)	***F*(2,37) = 6.05 p < 0.01** (**Elite vs**	***F*(1,37) = 13.58 p < 0.01**
	End	76.12 (41.99)	612.75 (614.77)	344.68(338.81)		
Comparison within groups		*F*(1,10) = 3.41 p = 0.095	***F*(1,13) = 7.53 p < 0.05**	***F*(1,14) = 10.86 p < 0.01**	Time*Group Interaction ***F*(2,37) = 3.51, p < 0.05**	
B	Start	77.47 (42.13)	79.73 (40.04)	100.68 (89.64)	*F*(2,37) = 0.77 p = 0.469	***F*(1,37) = 5.03 p < 0.05**
	End	114.50 (82.35)	412.02 (577.51)	319.10 (756.67)		
Comparison within groups		*F*(1,10) = 4.28 p = 0.065	*F*(1,13) = 4.50 p = 0.054	*F*(1,14) = 1.52 p = 0.238	Time*Group Interaction *F*(2,37) = 0.90, p = 0.416	

3.4 Cognitive Test –PenScreen (PS)

PS tasks of non-matching pairs with active distracters (NAC) and matching pairs with neutral distracters (MNC) tasks was found to be a promising screening tool for drivers who have malingering intent. The following tables described the significant differences between and within group comparison for NAC and MNC tasks.

Table 5. NAC mean RT: non-matching pair – active distracters

Trial	Time	Elite	Vulnerable	Malinger	Comparison by Group	Comparison by Time
A	Start	852.34 (273.70)	815.16 (145.65)	779.58 (95.19)	*F*(2,37) = 2.67 p = 0.083	***F*(1,37) = 7.53 p < 0.01**
	End	810.85 (204.10)	1160.71 439.69)	863.53 (248.06)		
Comparison within groups		*F*(1,10) = 0.90 p = 0.365	***F*(1,13) = 9.62 p < 0.01**	*F*(1,14) = 1.95 p = 0.185	**Time*Group *F*(2,37) = 5.73, p < 0.01**	
B	Start	790.43 (164.80)	827.28 (171.39)	744.82 (103.58)	*F*(2,37) = 1.38 p = 0.265	*F*(1,37) = 3.59 p = 0.066
	End	802.07 (176.05)	894.31 (204.76)	791.65 (158.41)		
Comparison within groups		*F*(1,10) = 0.52 p = 0.490	*F*(1,13) = 2.00 p = 0.181	*F*(1,14) = 1.68 p = 0.218	Time*Group *F*(2,37) = 0.51, p = 0.606	

Table 6. NAC SD: non-matching pair – active distracters

Trial	Time	Elite	Vulnerable	Malinger	Comparison by Group	Comparison by Time
A	Start	186.61 (125.86)	173.75 (86.46)	160.05 (53.03)	***F*(2,37) = 5.30 p < 0.01 (Elite & Malinger vs Vulnerable)**	***F*(1,37) = 20.51 p < 0.01**
	End	200.88 (131.28)	622.11 (440.57)	316.17 (212.09)		
Comparison within groups		*F*(1,10) = 0.16 p = 0.702	***F*(1,13) = 15.64 p < 0.01**	***F*(1,14) = 9.64 p < 0.01**	**Time*Group *F*(2,37) = 7.71,**	
B	Start	156.28 (106.62)	241.20 (275.08)	141.36 (64.60)	*F*(2,37) = 1.918 p = 0.162	*F*(1,37) = 2.55 p = 0.119
	End	179.57(85.85)	295.11 (139.59)	244.68(252.57)		
Comparison within groups		*F*(1,10) = 0.81 p = 0.390	*F*(1,13) = 0.55 p = 0.472	*F*(1,14) = 2.09 p = 0.172	Time*Group *F*(2,37) = 0.38, p = 0.688	

Table 7. MNC mean RT: non-matching pair – neutral distracters

Trial	Time	Elite	Vulnerable	Malinger	Comparison by Group	Comparison by Time
A	Start	720.85 (146.08)	697.14 (80.12)	693.74 (80.26)	*F*(2,37) = 1.15	***F*(1,37) = 5.66 p < 0.05**
	End	717.07 (146.88)	888.54 (314.99)	750.57 (184.96)		
Comparison within groups		*F*(1,10) = 0.041 p = 0.843	***F*(1,13) = 5.00 p < 0.05**	*F*(1,14) = 2.03 p = 0.176	Time*Group *F*(2,37) = 2.79,	
B	Start	705.00 (113.55)	736.38 (243.98)	688.93 (102.48)	*F*(2,37) = 0.69 p = 0.510	*F*(1,37) = 0.015 p = 0.902
	End	683.30 (106.69)	751.19 (138.15)	704.30 (90.06)		
Comparison within groups		*F*(1,10) = 2.48 p = 0.146	*F*(1,13) = 0.058 p = 0.813	*F*(1,14) = 0.88 p = 0.365	Time*Group *F*(2,37) = 0.26, p = 0.771	

Table 8. MNC SD: non-matching pair – neutral distracters

Trial	Time	Elite	Vulnerable	Malinger	Comparison by Group	Comparison by Time
A	Start	129.14 (59.11)	147.84 (63.08)	146.53 (110.15)	*F*(2,37) = 2.92 p = 0.067	***F*(1,37) = 19.36 p < 0.01**
	End	193.25 (128.98)	397.31 (270.73)	256.68 (169.27)		
Comparison within groups		*F*(1,10) = 3.95 p = 0.075	***F*(1,13) = 11.80 p < 0.01**	***F*(1,14) = 6.01 p < 0.05**	Time*Group *F*(2,37) = 3.00,	
B	Start	137.49 (53.77)	228.32 (402.05)	128.46 (58.52)	*F*(2,37) = 2.08 p = 0.140	*F*(1,37) = 1.08 p = 0.305
	End	141.26 (55.14)	288.02 (212.75)	177.65 (108.77)		
Comparison within groups		*F*(1,10) = 0.07 p = 0.803	*F*(1,13) = 0.38 p = 0.549	*F*(1,14) = 3.15 p = 0.098	Time*Group *F*(2,37) = 0.206, p = 0.815	

3.5 Monitoring System (Smart Eye Anti-Sleep, AS)

This is an off-the-shelf eye tracker (Smart Eye AB, Sweden) that can be mounted onto any vehicle to collect eyelid opening data at 60 times a second (60 Hz). It operates over a wide range of ambient lightings from dark to bright daylight with the capabilities to cancel spectacle reflections which may interfere with the tracking. The unit of measure is percentage PERCLOS, which stands for the proportion of time in a minute that the eyes are at least 80 percent closed (Wierwille et al., 1994)[20]. It reflects slow eyelid closures rather than blinks. This parameter is simple yet sensitive to driver fatigue making it a hot topic for research in driver fatigue for the past half a decade. With advance in technologies, PERCLOS can be derived real-time using eye tracking systems like AS. Even this, the AS system like others need to be validated with an Asian population where people generally have smaller eyes. The version of AS used in this study is tuned towards research where data can be logged and post analyzed for PERCLOS, allowing the researcher to fully understand the behaviour of data over time.

Thirty-three participant's data was analysed. Percentage of Eyelid Closure over a minute (PERCLOS) was statistically significant between 5 minutes at the start of driving (BP) and the point of 1-sec microsleep (P1) as shown in Fig.1. Traditional P-80 criteria where eyelid closure was defined as 80% eye closed yields 3.69% and 8.30% PERCLOS at BP and P1 respectively, and this difference was statistically significant ($F(1.49,44.65) = 7.8$, $p < 0.01$). The simpler but novel EO-7 criteria defined eyelid closure as 7 mm system eye opening reading (approximately 3mm actual eyelid opening corresponded to 2/3 eye closure). This EO-7 criteria yields 10.66% and 20.16% PERCLOS at BP and P1 respectively, and this difference was also statistically significant ($F(1.77, 53.05) = 12.58$, $p < 0.01$). EO-7 generated wider differences between alert and fatigued state and fewer tendencies for Type 1 error without the need for algorithms to determine baseline eyelid opening and to remove blink data.

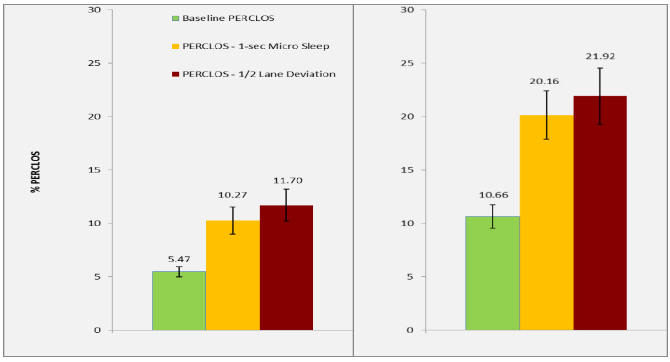


Fig. 1. Multiple comparisons between different data sets

The results and recommendation in this study will provide evidence for the customization of embedded AS suitable for Singapore population context. The analysis of data is steered towards how to make AS's PERCLOS measurement as hassle-free as possible for implementation in typical driving context.

4 Conclusions

The study had successfully derived screening, monitoring criteria and a prediction method for driver fatigue as part of risk management. It was proposed that PVT and PenScreen could be deployed as screening tools while Smart Eye Anti-Sleep PERCLOS was the recommended monitoring tool and using eye pupil tracking for fatigue prediction to reduce driving risk.

References

1. Brown, I.D.: Driver Fatigue. Human Factors 36, 298–314 (1994)
2. Oron-Gilad, T., Shinar, D.: Driver Fatigue Among Military Truck Drvers. Transportation Research, Part F (3), 195–209 (2000)
3. Barr, L., Popkin, S., Howarth, H.: An Evaluation of Emerging Driver Fatigue Detection Measures and Technologies. U.S. Department of Transportion - Federal Motor Carrier Safety Adminstration, Washington, USA (2009)
4. Dinges, D.F., Pack, F., Williams, K., et al.: Cumulative sleepiness, mood disturbances and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 h per night. Sleep 20, 267–271 (1997)
5. Tiplady, B.: The use of a mobile phone to administer a cognitive task. Poster presented at the Psychobiology Section of the British Psychological Society, Lake District (September 2004)
6. Edwards, D.J., Sirois, B., Dawson, T., Aguirre, A., et al.: Evaluation of fatigue
7. Management technologies using weighted feature matrix method", Proceeding of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Stevenson, Washington (2007)
8. LeDuc, P.A., Greig, J.L., Dumond, S.L.: Involuntary eye responses as measures of fatigue in US Army Apache aviators. Aviation, Space, and Environmental Medicine 76, C86–C91 (2005)
9. Lim, C.L., Yap, M.C.: Overall Final Report - D8: Optimal Shift Cycle Phase 2 - The effects of a single dose of caffeine in reversing the sleep-inducing effects of a single dose or melatonin or zolpidem. Technical Report, DMERI, DSO National Laboratories (2008)
10. Lin, S.T.: Ocular Motor Fatigue Induced by Prolonged Visual Display Terminal (VDT) Tasks. Honours Thesis, Department of Optometry & Vision Science, University of Melbourne, Australia (2007)
11. Rowland, L., Krichmar, J., Sing, H., Thomas, M., Thorne, D.: Pupil dynamics and eye movements as indicators of fatigue and sleepiness. Sleep Research 26, 626 (1997)
12. Russo, M., Thomas, M., Thorne, D., Sing, H., et al.: Oculomotor impairment during chronic partial sleep deprivation. Clinical Neurophysiology 114, 723–736 (2003)

13. Wilhelm, B., Giedke, H., Lüdtke, H., Bittner, E., et al.: Daytime variations in central nervous system activation measured by a pupillographic sleepiness test. Journal of Sleep Research 10, 1–7 (2001)
14. Fairclough, S.H., Graham, B.: Impairment of driving performance caused by sleep deprivation or alcohol: A comparative study. Human Factors 41(1), 118–128 (1999)
15. Larue, G.S., Rakotonirainy, A., Pettitt, A.: Driving performance impairments due to hypovigilance on monotonous roads. Accident Analysis & Prevention 43, 2037–2046 (2011)
16. Li, Z., et al.: Effects of acupuncture on heart rate variability in normal subjects under fatigue and non-fatigue state. Eur. J. Appl. Physiol. 94(5-6), 633–640 (2005)
17. Nam, K.C., Kwon, M.K., Kim, D.W.: Effects of posture and acute sleep deprivation on heart rate variability. Yonsei Med. J. 52(4), 569–573 (2011)
18. Wehrens, S.M.T., Hampton, S.M., Skene, D.J.: Heart rate variability and endothelial function after sleep deprivation and recovery sleep among male shift and non-shift workers. Scand. J. Work Environ. Health 38(2), 171–181 (2012)
19. Barr, L., Popkin, S., Howarth, H.: An Evaluation of Emerging Driver Fatigue Detection Measures and Technologies. U.S. Department of Transportion - Federal Motor Carrier Safety Adminstration, Washington, USA (2009)
20. Rigas, G., Goletsis, Y., Bougia, P., Fotiadis, D.I.: Towards driver's state recognition on real driving conditions. International Journal of Vehicular Technology, Article id 617210 (2011)
21. Wierwille, W.W., Ellsworth, L.A., Wreggit, S.S., Fairbanks, R.J., Kirn, C.L.: Research On Vehicle Based Driver Status/Performance Monioring: Development, Validation, And Refinement of Algorithms For Detection of Driver Drowsiness. Final Report: DOT HS 808 247, National Highway Traffic Safety Adminstration, Washington, USA (1994)
22. Henson, D., Emuh, T.: Monitoring vigilance during perimetry by using pupillography. Ophthalmology & Visual Science 51(7), 3540–3543 (2010)
23. Miró, E., Cano-Lozano, M., Buela-Casal, G.: Electrodermal activity during total sleep deprivation and its relationship with other activation and performance measures. Journal of Sleep Research 11(2), 105–112 (2002)
24. Pazderka-Robinson, H., Morrison, J., Flor-Henry, P.: Electrodermal dissociation of chronic fatigue and depression: evidence for distinct physiological mechanisms. International Journal of Psychophysiology 53, 171–182 (2004)

Novel Tools for Driving Fatigue Prediction: (1) Dry Eeg Sensor and (2) Eye Tracker

Frederick Tey[1], Sheng Tong Lin[1], Ying Ying Tan[1], Xiao Ping Li[2], Andrea Phillipou[3], and Larry Abel[3]

[1] Combat Protection & Performance Programme, DSO National Laboratories, 27 Medical Drive #09-00, Singapore 117510, Singapore
[2] Neuroengineering Lab, Department of Mechanical Engineering & Division of Bioengineering, National University of Singapore, Singapore
[3] Department of Optometry & Vision Sciences, University of Melbourne, Australia
tliankhe@dso.org.sg

Abstract. National Sleep Foundation's Sleep in America (2005) reported 60% of adult drivers driving a vehicle while feeling drowsy in the past year, and more than 37% have actually fallen asleep at the wheel [1]. This paper presented the findings of two novel fatigue prediction tools. The first study presents a 4-channel dry EEG under simulated driving being able to predict when the driver will develop microsleep in the next 10 minutes using only 3 minutes data of collected, with an accuracy of more than 80%. The second study uses an eye tracker to assess the percentage of time that the eyelids were closed (PERCLOS) as a potential marker for fatigue. Results showed that the average magnitude of oscillation (amount of pupil fluctuation), known as Coefficient Magnitude (CM), is generated from real-time wavelet analysis, has the potential to predict fatigue 8-12 minutes ahead with 84% accuracy ahead of compromised driving behavior.

Keywords: Fatigue, dry EEG, eye tracker, microsleep.

1 Introduction

Drowsiness/fatigue is a well known major risk factor for traffic accident. According to data from Australia, England, Finland, and other European nations, drowsy driving represents 10 to 30 percent of all crashes. It was also reported 60% of adult drivers saying they have driven a vehicle while feeling drowsy in the past year, and more than 37% have actually fallen asleep at the wheel [1]. Seriously, of those who have nodded off, 13% say they have done so at least once a month. These numbers may still underestimate the true frequency since it is difficult to assess driver sleepiness objectively [2]. Moreover, drivers themselves are sometimes unaware of sleepiness [3], resulting in the unreliability of subjective assessment.

Driving is a complex task involved numerous and varied brain functions such as attention, perception, memory, decision making. Thus, numerous parameters of given

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 618–627, 2013.

driving situation can affect these processes, complicating the study of driving fatigue [4]. In general, the causes of fatigue and drowsy driving could be classified into two broad categories: endogenous and exogenous factors [5].

The endogenous factors related to the physiological processes underlying alertness or wakefulness, including the circadian variations associated with time of day, the fatigue generated with the duration of the task and sleep-related problems. Long hours spent driving, referred to as the time-on-task effect, are known to produce fatigue and a deterioration of driving performance, although the degradation can occur during the very early stages of a driving or vigilance task. It is observed that 60% of fatal sleep-related accidents in Finland occurred within the first hour of driving [6]. The time-of-day effect is another major factor that accounts for fatigue. From a perspective of sleep-wake regulation, the time-of-day effect is driven by two key neurobiological processes: a homeostatic process producing a progressive sleep drive over time awake and a circadian process producing an opponent wake drive as a function of time of day [7]. The homeostatic process, known as the sleep "homeostat", balances time spent awake and time spent asleep. While the circadian, which originates in the biological clock in the suprachiasmatic nuclei of the hypothalamus [8], process keeps track of time of day and night. The homeostatic and circadian processes interact to produce a combined influence on driving performance. Park et al. (2005) showed that a significant portion of sleep-related accidents happened during the early morning (2am-6am) and during the afternoon period (2pm-4pm) [10]. Sleep-related factors such as sleep deficit and sleep deprivation also increase accident risk. Some studies have highlighted that even for short duration of driving (1-2h) and moderate time awake (8h) sleep restriction still can impair driving [11].

The exogenous factors could be the characteristics of road geometry and roadside environment, or other factors that define the driving task. These factors can have an impact on driving performance by affecting attention, alertness and information processing [5],[12]. It is acknowledged that highway night drivers are particularly vulnerable to sleep-related accidents [12] due to the road geometry and roadside environment remain unchanged or highly repetitive. An under-demanding monotonous road environment with low traffic density can result in feelings of boredom and drowsiness coupled with loss of interest of performing the task at hand, eventually lead to sleep. Performance deterioration, which is induced by under load, may be as important as what is observed during over-demanding crowded urban expressway situations, when arousal is raised to a point where the driving performance is negatively affected.

Endogenous and exogenous factors interact continuously and it is their joint influence that determines alertness and vigilance at any given point during driving.

Extensive research has been conducted to develop systems for monitoring the driver's level of sleepiness using different techniques, such as measures of brain wave, heart rate, electrocardiogram, respiration and eye tracker.

2 Development of Dry EEG Sensor Headsets and Signal Processing Software for Driving Fatigue Prediction

2.1 Background

Electroencephalograph (EEG) has been acclaimed as one of the most predictive and reliable measurements since it directly and immediately reflects human brain state or brain activity. There has been considerable evidence showing the possibility of EEG-based detection of early stages of sleepiness. In this study, a dry sensor based mobile EEG recording system is developed offering prolonged high-quality EEG recording under real-life driving conditions. In addition, a long-term prediction of sleep onset based on the biological mechanism of sleep which is more reliable and accurate, is proposed to allow predicting unintentional sleep onset long before the eventual sleep onset.

Sleep can be induced by two processes: passive process due to closure of cerebral gates (brain deafferentation) and active process promoted by inhibitory mechanisms arising in some cerebral areas. The passive process of sleep initiation depends on the regulation of homeostatic. The initiation of sleep is a consequence of the dampened activities in our wake-promoting brain systems. In addition, the absence of a steady excitatory bombardment may produce disfacilitation in some brain structures, eventually followed by rebound cellular excitation that would set in motion a series of structures which may promote sleep through active inhibitory processes.

2.2 Methods

The experiment was separated into two parts. In the first part, EEG data were acquired using traditional wet EEG sensors, while innovative dry EEG sensors were used in the second part.

Participants. Sixteen young healthy men, range 21–25 years, ten in the first part and six in the second part, participated in the study. All the participants were recruited from the National University of Singapore and Nanyang Technological University, had normal or corrected to normal vision, reported no history of neurological problems and were right-handed. Subjects were required to keep a sleep diary one week prior to the experiment to ensure that they had at least 7 hours of continuous sleeping time and regular sleeping hours (going to bed no later than 1 am and waking up by 9 am). Informed consent was obtained from all participants in accordance with the guidelines and approval of the National University of Singapore Institutional Review Board. The subjects will be reimbursed for each trial for their participation.

Procedure. In the first part, all the subjects were tested in the evening from 8pm to 2am next day which is associated with the peak for nocturnal sleep-related accidents. Participants were asked to have dinner early and come to the laboratory at 7.30 pm.

The experiment lasted for six hours continuously. Subjects were asked to sit comfortably and avoid unnecessary movements such as singing, mumbling, or talking during the experiment. In the second part, all the settings are kept consistent, only the time of the experiment was changed. The subjects were tested from 10 am to 2 pm, which is considered as mid-day dip period.

Simulated Driving Task. Simulated driving tests consisted of a video clip showing moving road images of monotonous highways mostly free of vehicles shown from the perspective of a driver while operating a car. Subjects were instructed to watch the road at all times and response by pressing keyboard buttons (right arrow and left Ctrl) when the car changed lanes. Simulated driving lasted for six hours. The driver's attempts at maintaining vigilance were dependent on their own determination to stay awake.

Data Acquisition. EEG data was acquired using an ANT amplifier (Advanced Neuro Technology, Enschede, Netherlands) connected to an electrode cap. Wet and dry electrodes were mounted on the cap based on the International 10-20 electrode placement system. All channels were referenced to the link of the left and right mastoids, and grounded with channel AFz. Input impedances of all channels were kept below 20 kΩ for all experimental sessions. Data was sampled at 250 Hz and recorded using ASA-lab software from Advanced Neuro Technology.

2.3 Results

As shown in the Table 1, the algorithm could predict the 2-second microsleep around 10 minutes ahead (9.75±2.76) with 80% accuracy. The time ahead for subject 5 (1 minute), subject 9 (3 minutes), and subject 11 (4 minutes) is less than 5 minutes, but can still be useful for the warning purpose. The prediction accuracy and time ahead for 1-second microsleep was lower compared to 2-second microsleep. The system could only predict accurately for 62% of the subjects with 3.5±1.26 minutes ahead.

The final portable system only requires 6-channels EEG headset, 4 channels (Fz, Cz, Pz, Oz) for brain signals and 2 reference channels (Fig 1) to provide real-time countermeasure device for preventing sleepiness or unintentional sleep onset related accidents.

Spindles appear during early states of sleep in the 9-15 Hz band of EEG signals, is the biological marker for the release of GABA and spindle properties reflect the accumulation of GABA in the brain. Therefore, a simple measure of spindles can provide the level of fatigue state at given time. The algorithm for spindle detection achieved 91.71% accuracy with traditional wet EEG data and a lower accuracy of 89.31% was achieved with dry EEG sensors possibly due to the interferences of noise after a few hours of continuously recording.

Table 1. Fatigue prediction results

Subject No.	2-s Microsleep	Prediction Time	Time Ahead	1-s Microsleep	Prediction Time	Time Ahead
1	65	51	14	55	50	5
2	42	34	8	36	33	3
3	28	20	8	15	Fail	
4	18	12	6	16	12	4
5	25	24		23	Fail	
6	29	16	13	18	16	2
7	20	14	6	17	14	3
8	55	45	10	29	Fail	
9	13	10		11	10	
10	214	204	10	86	83	3
11	49	45		47	44	3
12	52	45	7	46	45	
13	45	34	11	25	Fail	
14	121	110	11	78	74	4
15	non			249	243	6
16	40	27	13	29	27	2
		Mean	9.75		Mean	3.5
		SD	2.76		SD	1.26
**All units are in minute*						

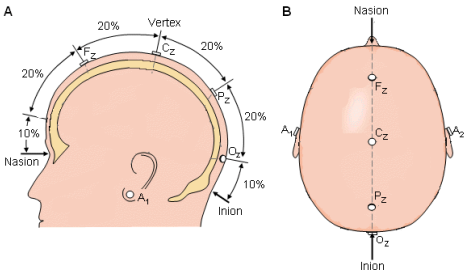


Fig. 1. Electrodes diagram. (The 4 electrodes Fz, Cz, Pz, Oz is placed based on 10-20 scheme. Reference electrodes A_1 and A_2 should be placed on "inactive" zones such as ear lobes or mastoids).

To this end, a dry EEG sensor offering prolonged high-quality EEG recording under real-life driving conditions was developed. It is capable of providing prediction of sleep 10 minutes ahead of time with an accuracy of 80% based on the biological

mechanism of sleep instead of the statistic machine learning and pattern classification methods. This is an important finding since it simplifies the design of EEG headset, make it really portable and practical. In addition, reducing number of channels also means less calculation time and less required memory for computation.

3 Using Wavelet Analysis for Pupil Oscillation to Predict Fatigue Onset

3.1 Background

In 1993, Irene Loewenfeld, the famous scientist in pupilometry, published her comment [14]: "(Fatigue waves) are involuntary and unconscious, so that they cannot be produced deliberately; and best of all, running records can be obtained without touching the subject. These show the slightest fluctuations from one moment to the next, from day to day, from week to week, and over longer periods." Since then, fatigue-induced pupil oscillations below 0.8 Hz have attracted the attention of numerous sleep experts [15-20].

The basis of this oscillation lies in the activity of the autonomic nervous system in the human body which comprises of sympathetic and parasympathetic nervous system. Parasympathetic system is responsible for pupil constriction. In normal alert state, sympathetic system plays the role of inhibiting parasympathetic system, thus allowing stable dilatation of pupil for active vision and perception. During sleepy or fatigue state, the inhibitory function of the sympathetic system is impaired. Hence, the infrequent firing of inhibitory signal on the parasympathetic system causes a drift in this stability of pupil size which is manifested as pupillary oscillation.

In 1998, Lüdtke & colleagues formalized a standard called Pupillary Unrest Index (PUI), to measure fatigue in a dark room setting [17]. Due to its limitation and inconvenience as a real-time fatigue monitoring method, other scientist are looking at low frequency power of pupil fluctuation using frequency power spectrum [15,19] and frequency wavelet transformation [20]. Henson & Emuh (2010) found that below 0.8 Hz, the signal amplitude for pupil oscillation increases with fatigue significantly. This study attempts to validate the work of Henson & Emuh (2010) by applying their techniques on pupil data collected from real driving and indoor.

3.2 Data

While 44 participants' data were collected, only 32 of them were analyzed. The rest of the participants are excluded due to wrong parameter setting (1), more than 40% data loss (5), microsleep less than 12 mins into 1st hour driving (3), corrupted video data (1) and strong drivers without microsleep (2).

3.3 Method of Analysis

Raw data from driving were processed into dependent variables (outcome or measuring variables) using Matlab 2012. Firstly, eye blink and other noise artifacts in

the pupil size data were smoothen out using interpolation and median filtering algorithms available from Matlab before broken down into 4-minutes segments. Then, each of these segment undergone wavelet transformation computation using the "reverse biorthogonal 3.7" method (used by Henson & Emuh, 2010), also available from Matlab's Wavelet Toolbox. Signal amplitudes (or strength) in arbitrary units known as Coefficient Magnitudes (CM) for 60 low pupil oscillation frequencies ranging from 0.012 Hz to 0.727 Hz were produced from each of these transformed segments. Only CMs for frequencies 0.242 Hz, 0.104 Hz, 0.03 Hz and 0.02 Hz were selected as the dependent variables for statistical analysis. This method of deriving CMs from pupil fatigue wavelet was validated in simulated driving in Dr Larry Abel's lab. Dr Abel found that CMs derived this way is relatively "undisturbed" by ambient lighting changes like that of the outdoor driving. For the purpose of fatigue prediction, selected frequencies' CMs were picked for statistical analysis from the following Fatigue Segments into driving.

1) BP - Baseline Fatigue Segment 8-12 minutes into Baseline driving
2) MS -12mins - Fatigue Segment 8-12 minutes after MS
3) MS -8mins - Fatigue Segment 4-8 minutes after MS
4) MS -4mins - Fatigue Segment 0-4 minutes after MS
5) MS - Fatigue Segment where the 1-second microsleep (sleep eye closure for more than 1 second) occurs within.
6) End - 4-8 minutes just before end of driving, completed 4 hours or up to the hour where lane deviation occurs

Based on the distribution of 1-second microsleep time during the trial, the participants were also divided into the following groups:

1) Tolerance Group A (n = 15) - 1-second microsleep in less than 30 minutes into driving
2) Tolerance Group B (n = 9) - 1-second microsleep between 30 to 60 minutes into driving
3) Tolerance Group C (n = 8) - 1-second microsleep after 60 minutes into driving.

All statistical analysis was done using IBM SPSS Version 20. For each frequency, a mixed design ANOVA was used to test for significance differences between Tolerance Group, Fatigue Segments and their interaction. Indoor data were also processed and analyzed in the same manner except that Fatigue Segments are defined as Indoor baseline (data taken in the morning before game driving) and Indoor End (data taken in the evening after monotonous game before participants were released).

3.4 Results

For frequencies 0.242 Hz and 0.104 Hz, their MS-12mins shown strong statistical differences when compared to BP and MS, thus, a high reliability for predicting MS; in another words, a CM value that lies between BP's and MS's yet clearly defines itself away from the two. Hence, by carefully selecting the CM readings based on MS-12mins' confidence interval, it is possible to set a fatigue warning threshold (See Fig 2).

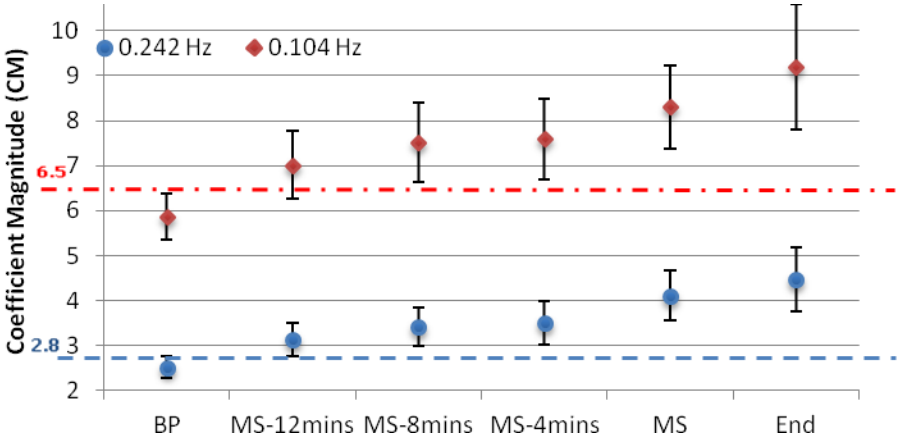


Fig. 2. Pupil oscillations for 4 oscillation frequencies for indoor data

Note that two thresholds were selected (Fig 2, red dotted line and blue dotted line), CM 2.8 for 0.242 Hz and CM 6.5 for 0.104 Hz. The selections were made based on two criteria, that is, the threshold must not be below the upper bound of BP's 95% CI and must be as close as possible to the lower bound of MS-12mins' 95% CI. The rational is to avoid false alarm by setting it sufficiently above BP level but conservative enough to signal MS as much ahead of time where possible. Since the MS-12mins segment was used, these thresholds predicted the first 1-second microsleep, 12-8 minutes before its occurrence. When these thresholds were applied back to the 32 participants' data, 78% and 72% prediction accuracy were achieved for the 0.242 Hz and 0.104 Hz frequencies' thresholds respectively, and 84% when both thresholds were applied (those participant's data that could not be predicted by the 0.242 Hz threshold were re-checked with the 0.104 Hz threshold).

If the CM 2.8 and 6.5 thresholds were applied back to the 22 participants, we found that some participants can be notified of their 1-sec microsleep as early as 28 minutes ahead. Note that 86.3% of the 22 participants can be notified of their 1-second microsleep at least 12 minutes before their 1-second microsleep. This is similar to the 84% found with the original pool of 32 participant's data. Although the thresholds can predict fatigue much earlier for certain people due to individual differences in CM manifestation, the project team would like to maintain a conservative definition of their suggested thresholds: CM 2.8 and CM 6.5 only predicts onset of 1-second microsleep 12 to 8 minutes ahead. This is due to the low number of participants in this part of the analysis which do not give enough data confidence for more refine conclusion on prediction time.

4 Conclusion

This paper has successfully demonstrated two novel fatigue tools. Using eye-tracker, the research team has identified the pupil behaviour threshold, in terms of low frequencies oscillation signal strength, which could predict onset of 1-second

microsleep 8-12 minutes ahead of its onset with 84% accuracy in the study population. In addition, a separate study has successfully identified the EEG brain signal which indicates the onset of 1-second microsleep 5 minutes ahead at 62% accuracy and 2-second microsleep 10 minutes ahead at 80% accuracy. These are important steps towards a real-time countermeasure device for preventing sleepiness or unintentional sleep onset related accidents.

References

1. Summary Findings of the Sleep in America poll (2005), http://www.sleepfoundation.org/sites/default/files/2005_summary_of_findings.pdf
2. Rosekind, M.R.: Underestimating the societal costs of impaired alertness: safety, health and productivity risks. Sleep Med. 25 (suppl. 1), S21–S25 (2005)
3. Horne, J.A., Baulk, S.D.: Awareness of sleepiness when driving. Psychophysiology 41, 161–165 (2004)
4. Smiley, A.: Fatigue management: lessons from research. In: Hartley, L. (ed.) Managing Fatigue in Transportation, pp. 1–23. Elsevier, Oxford (1998)
5. Thiffault, P., Bergeron, J.: Monotony of road environment and driver fatigue: a simulator study. Accident Analysis & Prevention 35, 381–391 (2003)
6. Summala, H., Mikkola, T.: Fatal accidents among car and truck drivers: effects of fatigue, age and alcohol consumption, age and alcohol consumption. Human Factors 36, 315–326 (1994)
7. Van Dongen, H.P.A., Dinges, D.F.: Sleep, circadian rythms, and psychomotorvigilance. Clinics in Sport Medicine 24, 237–249 (2005)
8. Moore, R.Y.: Organization of the mammalian circadian system. In: Circadian Clocks and Their Adjustments, pp. 88–106 (1995)
9. Pack, A.I., Pack, A.M., Rodgman, E., Cucchiara, A., Dinges, D.F., Schwab, C.W.: Characteristics of crashes attributed to the driver having fallen asleep. Accident Anal. Prevention 27, 769–775 (1995)
10. Park, S.-W., Mukherjee, A., Gross, F., Jovanis, P.P.: Safety Implications of Multi-day. Driving Schedules for Truck Drivers: Comparison of Field Experiments and Crash Data Analysis. Transportation Research Board Annual Meeting CD, Journal of the Transportation Research Board (2005)
11. Philip, P., Sagaspe, P., Moore, N., Taillard, J., Charles, A., Guilleminault, C., Bioulac, B.: Fatigue, sleep restriction and driving performance. Accident Analysis & Prevention 37, 473–478 (2005)
12. Larue, G.S., Rakotonirainy, A., Pettitt, A.N.: Driving performance impairments due to hypovigilance on monotonous roads. Accident Analysis & Prevention 43, 2037–2046 (2011)
13. Åkerstedt, T., Czeisler, C.A., Dinges, D.F., Horne, J.A.: Accidents and sleepiness: a consensus statement from the International Conference on Work Hours, Sleepiness and Accidents, Stockholm, September 8-10, pp. 8–10 (1994); J. Sleep Res. 3, 195
14. Loewenfeld, I.: The Pupil. Anatomy, physiology and clinical applications. Wayne State University Press, Detroit (1993)
15. McLaren, J., Erie, J., Brubaker, R.: Computerized analysis of pupillograms in studies of alertness. Investigative Ophthalmology & Visual Science 33(3), 671–676 (1992)

16. Nishiyama, J., Tanida, K., Kusumi, M., Hirata, Y.: The pupil as a possible premonitor of drowsiness. IEEE, 1586–1589 (2007)
17. Lüdtke, H., Wilhelm, B., Adler, M., Schaeffel, F., Wilhelm, H.: Mathematical procedures in data recording and processing of pupillary fatigue waves. Vision Research 38, 2889–2896 (1998)
18. Wilhelm, H., Lüdtke, H., Wilhelm, B.: Pupillographic sleepiness testing in hypersomniacs and normals. Graefe's Archive for Clinical and Experimental Ophthalmology 236, 725–729 (1998)
19. Nakayama, M., Yamamoto, K., Kobayashi, F.: Estimation of sleepiness using frequency components of pupillary response. IEEE, 357–360 (2008)
20. Henson, D., Emuh, T.: Monitoring vigilance during perimetry by using pupillography. Investigative Ophthalmology & Visual Science 51(7), 3540–3543 (2010)

Quantifying Resilience to Enhance Individualized Training

Brent Winslow[1], Meredith Carroll[1], David Jones[1], Frank Hannigan[1], Kelly Hale[1], Kay Stanney[1], and Peter Squire[2]

[1] Design Interactive, Inc., 1221 East Broadway St., Oviedo, FL 32765
[2] Office of Naval Research, 875 North Randolph St., Arlington, VA 22217
brent.winslow@designinteractive.net

Abstract. Resilience is the human ability to adapt in the face of tragedy, trauma, adversity, hardship, and ongoing life stressors. To date, experimental reports on this subject have focused on long-term trajectories (weeks to months) of resilience, with little or no focus on whether significant changes to resilience could be achieved by short-term interventions. Currently, an individual's resilience is defined either by self-report or by behavioral changes such as the development of depression, post-traumatic stress disorder, or suicide. We propose that the quantification of an individual's physiological and behavioral response to stress under controlled conditions is an indication of the individual's level of resilience. To address such real-time resilience, we propose the first in a series of studies to evaluate real-time human resilience by exposing participants to controlled stressors while assessing the stress response. Activation of the hypothalamus-pituitary-adrenal cortex axis and sympathetic branch of the autonomic nervous system via monitoring of the pupil constriction, heart and respiration rate, muscle tonicity, salivary cortisol, and electrodermal activity will be assessed. Stress exposure will consist of virtual stressors presented using Virtual Battlespace 2 software-based scenarios, such as noise exposure, time pressure, and emotion-induction tasks, as well as external stressors such as socio-evaluative stress via the Trier social stress task, while evaluating decision-making and performance. The relationship between performance and the physiological stress response will be quantified, including the creation of a series of stress-performance trajectories based upon individual differences. Such an analysis is similar to probing for resilience in material testing, in which a load is applied to a candidate material, and the resulting forces and observable changes in dimension are quantified and reported via stress-strain curves. Ongoing studies will examine how this resilience measure may be integrated into a closed-loop training system to provide appropriate coping strategies to optimize resilience training. Such training programs, which take into account individual perceptions of stressors and physiological responses, are expected to be effective in helping trainees develop resiliency during high-stress operations.

Keywords: Resilience, Stress, Adaptation, Training, Autonomic Nervous System.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 628–636, 2013.

1 Introduction

Resilience is the human ability to adapt in the face of tragedy, trauma, adversity, hardship, and ongoing life stressors as defined by the corresponding MeSH term introduced in 2009 for use in indexing articles in PubMed [1]. As such, resilience includes both the concept of adaptation to stress, including maintenance of performance, and physiological bounce back, consisting of both the intensity of the response and the time to return to baseline [2]. Studies investigating resilience have increased rapidly in the past several years (see Figure 3F), many of which have focused on military-relevant aspects of resilience, including rates of post-traumatic stress disorder (PTSD) [3] and suicide, which have been increasing in the wake of the conflicts in Iraq and Afghanistan [4]. Other studies have focused on factors that augment or diminish resilience in individuals [5, 6]. To date, such reports have focused on long-term trajectories (weeks to months) of resilience, with little or no focus on whether significant changes to real-time resilience (i.e., in the moment, the ability to shut down counterproductive thinking to enable greater task-centric concentration and focus during high-stress operations) could be achieved by short-term interventions or training scenarios [7, 8]. Currently, an individual's resilience is qualitatively defined either by one of a number of self-report scales [9-11], or quantified by behavioral changes such as the development of major depressive disorder (MDD), PTSD or suicide. Self-report scales suffer from problems with exaggeration or under-reporting [12]. There is a great need to identify individuals at risk for problems with resilience prior to the development of MDD, PTSD, or suicide attempt.

The challenge in obtaining an objective measure of real-time resilience may be addressed by quantifying an individual's physiological and behavioral response to stress under controlled conditions. In general following exposure to significant stressors, behavioral performance including fine motor performance [13], attention [14], and cognitive function [15] decrease due to biological and neural mechanisms [16, 17]. When an individual encounters a significant stressor, the sympathetic or "fight or flight" division of the autonomic nervous system (ANS) increases in activity, resulting in neurotransmitter release and subsequent physiological effects on multiple organ systems. Among a myriad of effects, heart rate increases, pupils dilate, blood vessels constrict, and sweat glands become active. Severe stress also strongly activates the HPA axis [16, 17], consisting of the hypothalamus, pituitary, and adrenal glands, resulting in spikes of stress related hormones including cortisol. Cortisol activity is associated with reduced inflammation and immunity, muscle and fat loss and conversion of glucagon to glucose, which function to provide energy during stress or fasting conditions. Cortisol release is also governed by negative feedback loops in which free cortisol binds to glucocorticoid receptors (GCRs) in the anterior pituitary gland and the hypothalamus to decrease release of cortisol precursors. Individuals with clinical stress disorders have been found to have multiple epigenetic effects including decreased levels of GCRs in various brain areas [16, 17], so cortisol remains active for longer periods of time in such individuals. Stress-related ANS and HPA cascades depend on an individual's perception of the stressor, including its perceived novelty, controllability, and predictability, with those stressors or situations perceived as a

"threat" causing higher physiological stress responses than those seen in a more productive way as a "challenge" one has to contend with [18].

The quantification of load application and subsequent deformations has been employed in physics and engineering for many years to define material resilience – the capacity of materials to return to their initial shape following exposure to external forces, often modeled in stress-strain curves [19]. In such analyses, resilience can be calculated from the area under the curve in the linear zone, the region in which a material can return to its initial dimensions without permanent deformation. Material resilience is commonly improved by methods such as strain hardening, in which materials are cyclically deformed up to their elastic limit, ultimately resulting in a stronger material due to reorganization of intermolecular forces [20]. A similar phenomenon may be active in individuals in whom stress exposure increases resilience [21, 22]. In order to evaluate real-time human resilience, we propose the first in a series of experiments to expose participants to controlled stressors while monitoring activation of the HPA and sympathetic branch of the ANS via measurement of pupil constriction, heart and respiration rate, muscle tonicity, salivary cortisol, and EDA. The degree of change in physiological response along with behavioral performance measures will be evaluated to determine an individual's real-time resilience.

2 Experimental Methods

The proposed experiment will consist of a within subjects repeated measures design. All participants will perform three types of tasks: 1) small unit leader decision making scenarios in Virtual Battlespace 2 (VBS2, Bohemia, Orlando FL) in the absence of stressors, 2) small unit leader decision making scenarios in VBS2 in the presence of simulation-based stressors. 3) Socio-evaluative stress using a modified Trier Social Stress Task (TSST), or a stress-free control version. A total of 60 participants will be recruited, with 30 going through each version of the TSST. The experiment is designed to last approximately 2.5 hours, and the associated time-line is shown in Figure 1.

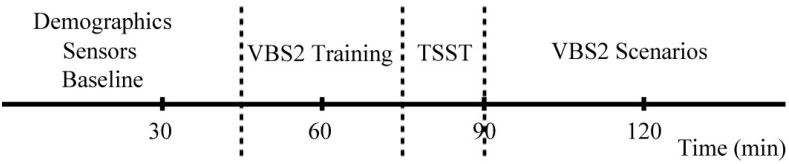


Fig. 1. Proposed experiment timeline

Measures of both ANS activity, as well as HPA reactivity will be collected throughout the TSST and VBS2 scenarios. A BioNomadix sensor suite (Biopac Systems, Goleta, CA) will be used to gather physiological data non-invasively from multiple body sites and locations (Figure 2). Electrocardiogram (ECG) sensors will be applied to the chest in a 3 lead configuration to record the ECG, and a respiration strap will be placed across the chest to gather breathing intensity and kinetics. Electromyogram (EMG) sensors will be placed on the trapezius muscle to gather electrical

activation and tension of this muscle group. Electrodermal activity (EDA) sensors will be placed on the fingers of the non-dominant hand to gather electrodermal responses (EDR), and a pulse plethysmography (PPG) sensor will be placed on the thumb to gather pulse. Skin temperature will also be monitored. All of these sensors stream data in synchrony and will be analyzed to identify the most robust classifiers of stress reactions. Heart rate variability (HRV) will be calculated from the ECG data in the time domain, by calculating the R-R interval standard deviation in 5 minute intervals (SDNN) [23]. Pulse transit time (PTT), which is the time it takes the pulse waveform to propagate from the heart to the periphery, and is indicative of sympathetic activation, will be calculated from the ECG and PPG at the thumb. In addition, vagal tone and respiratory sinus arrhythmia (RSA) will be derived from the ECG and respiration unit. Cortisol levels prior to and following exposure to stress will be captured via saliva samples, which will be tested for cortisol levels offline.

Fig. 2. Bionomadix sensor suite consisting of 3-lead ECG, respiration strap, EDA and PPG attached to the fingers of the non-dominant hand, as well as EMG sensors on the trapezius

The TSST is used as a positive control for stress and is divided into three 5-minute components [24]. Stress induction begins with the participant being taken into a room where a panel of 2-3 judges, described as being trained in public speaking, dressed in white lab coats, along with a video camera and audio recorder will be waiting. The first 5-minute component is the anticipatory stress phase, during which the judges ask the participant to prepare a 5-minute oral presentation describing why he or she is the best candidate for their ideal job. During the 5-minute presentation component, the judges will observe the participant without comment, and with neutral expressions throughout. At any point if the participant stops or does not use the entire five minutes, the judges will prompt him or her to continue. The presentation will be immediately followed by a mental arithmetic component, during which the participant is asked to verbally count backwards from 1,022 in steps of 13. If a mistake is made, the judges will prompt the participant to start again from 1,022. This component will last

for five minutes and is followed by a recovery period. The control group will receive another version of the TSST [25], which contains the same factors except for the psychosocially stressful components, including the socio-evaluative threat and uncontrollability. In this group, the first 5-minute component consists of reading a popular scientific text after being told that reading performance will not be evaluated. Participants will then be asked to read out the text in a low voice for five minutes, and finally to enumerate a series of numbers in increments of five in a low voice for another five minutes (e.g., 5, 10, 15, etc.). Two-three experimenters will be present during these sessions but they will neither wear white laboratory coats nor interrupt the participants. In addition, they will not observe or evaluate the participants nor will they ask any questions. There also will be no video-cameras present.

Scenarios will be created in VBS2 to simulate a tactical military environment with specific mission objectives, time requirements, and consequences depending on the course of action a participant pursues. Several scenarios will be designed to produce low levels of stress, such as following a person of interest throughout a town, or higher levels of stress such as clandestine demolition. Simulation-based stressors will be integrated into the scenarios, such as limited visual perception (night missions), sudden noise exposure, equipment failures, and receiving enemy fire, as well as cognitive tasks (e.g., time pressure) and emotion induction procedures, including dead combatants, soldiers and civilians [26]. VBS2 scenarios will be presented on a PC running on a Pentium i5 quad core processor, with 8 GB of ram and a high-end graphics card. Participant performance in scenario will be quantified by a number of process and outcome measures, including observation of high priority areas of interest via eye tracking, verbal reports, decision accuracy, and reaction time.

Non-contact sensors will gather affective and cognitive data. Facial expressions and verbalizations of participants will be recorded via webcam during experimental procedures to analyze affective state and communication offline. Visual fixations will be detected during scenarios by eye tracking using the easyGaze® eye tracker and gazeWare® software (Design Interactive, Oviedo, FL).

3 Expected Results

Expected stress response and performance results are shown graphically, compared to material tensile testing, in Figure 3. As described previously, material tensile testing consists of applying a load to a material of known dimensions, and measuring both the force applied (Figure 3A), and observing the change in length (Figure 3C) over time. Such data are then used to define the stress (σ) as Force/Area, and strain (ε) as Δ length / initial length, plotted against one another in a stress-strain curve. Expected results include changes in physiological responses to stressors over time, in which an individual's stress response is expected to vary with the stressor applied (Figure 3B). In addition, performance is expected to decrease at high levels of physiological stress (Figure 3D). An effort will be made to use similar logic behind material testing to quantify resilience, such that observable performance (the human analogue to strain)

is plotted against measureable physiological reactivity (the human analogue to stress; Δ physiological response / baseline physiological response), as shown in Figure 3F. It is expected that certain physiological measurements and calculations, such as electrodermal activity and pulse transit time, will be more indicative of stress response and resilience. It is also expected that different individuals will follow various different curves [7], and taken together, the data will allow for devising a means to model real-time human resilience prior to deployment.

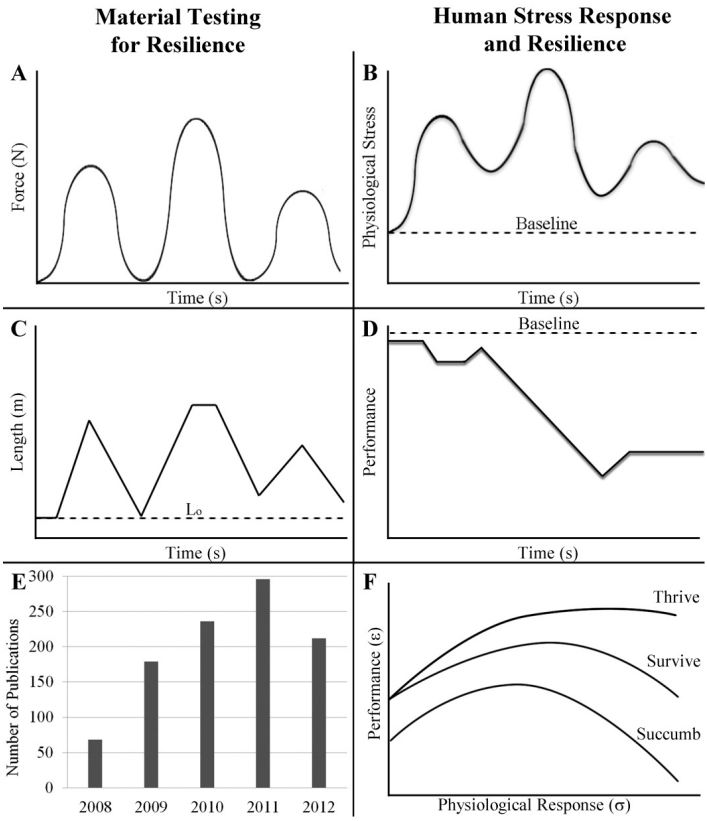


Fig. 3. Material testing for resilience compared to human stress testing. [A] In material testing, tension is applied to a material and the resulting forces are quantified over time, as well as changes to length [C]. Such data are then plotted in a stress (Force/unit area) vs. strain (length/initial length) graph, and the linear region gives a measure of material resilience. In human stress testing, psychological stressors are introduced, and the resulting physiological stress response [B], as well as changes to task performance [D] are quantified over time. A similar plot is used [F] to determine performance vs. physiological response and real-time resilience. Resilience articles have been steadily increasing, as shown by the Mesh search for the corresponding term [E].

4 Discussion

In this article, we have proposed a novel method to quantify human real-time resilience, by using similar methods and logic as in quantifying material resilience. Similar to material testing, a (human) system is challenged by introducing stress, and measure the response both in terms of observable performance (strain) and measurable physiological changes (stress). As one should expect from different materials which exhibit a range of intermolecular forces, it is expected that differences in the human stress strain curves will be found due to individual differences in real-time resilience [27]. However, unlike material testing, in which characterized materials exhibit relatively low variability, the human system may not yield consistent, linear, characterizable patterns within such an analysis due to higher variability. In addition, little variation in material characteristics or response is seen day-to-day in material testing, whereas such factors as exposure to life stressors [28] or time of day [29] may affect human real-time resilience. These factors will need to be addressed in the measure of real-time resilience.

Long-term changes to resilience have been shown to affect protein expressions and neuronal pathways. Peres et al. [8] showed that one month of psychotherapy given to police officers with PTSD significantly improved symptoms, but also increased prefrontal cortex (PFC) and decreased amygdala activity during traumatic recall. The amygdala – PFC circuit has been shown to be predictive of pathological stress reactions [30]. Another recent clinical trial has shown that a 12 week resilience-oriented treatment for PTSD increased clinical scoring on nearly every self-report test for resilience available [31]. However, the mechanisms underlying more rapid approaches to building real-time resilience have not yet been defined.

The approach taken in this study represents an endophenotype analysis of resilience, which was first described as an intermediate between genes and a disease state, and has proven useful in suicide analysis [32]. Such analyses have shown that suicide traits are measurable, but generally unobservable to the unaided eye. Similar to a cardiovascular stress test, the system must be challenged in order for traits to be expressed and observed. Along those same lines, the current work seeks to cause a physiological stress response in order to measure underlying changes in real-time resilience, which would otherwise be unobservable without such a challenge. Such an analysis may define at-risk individuals prior to deployment.

Future studies will examine how a real-time resilience measure may be integrated into closed-loop training systems to provide appropriate coping strategies at the right time to optimize resilience training within an operational context. Such adaptive training programs, which take into account individual perceptions of stressors and physiological responses, are expected to be effective in helping trainees develop resiliency during high-stress operations, since there is a wide heterogeneity in individual's responses to trauma and adversity. Future studies will also take advantage of stress-hardening techniques that are used in materials science to increase the strength of materials for specific applications by applying cyclic strain, without heating or modifying the materials, resulting in changes to material resilience. The human analogue

would be any number of stress-reduction techniques, including stress exposure and training such as described in this article, or other methods including psychotherapy [8], techniques of mental preparation [23], biofeedback [33], or appraisal [34].

References

1. Pubmed A service of the US National Library of Medicine and the National Institutes of Health (2009), http://www.ncbi.nlm.nih.gov/pubmed/
2. Carroll, M., et al.: Framework for training adaptable and stress-resilient decision making. In: Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Annual Meeting, Orlando, FL (2012)
3. Baker, D.G., et al.: Predictors of risk and resilience for posttraumatic stress disorder among ground combat Marines: methods of the Marine Resiliency Study. Prev. Chronic. Dis. 9, E97 (2012)
4. Milliken, C.S., Auchterlonie, J.L., Hoge, C.W.: Longitudinal assessment of mental health problems among active and reserve component soldiers returning from the Iraq war. JAMA 298(18), 2141–2148 (2007)
5. Obradovic, J.: How can the study of physiological reactivity contribute to our understanding of adversity and resilience processes in development? Dev. Psychopathol. 24(2), 371–387 (2012)
6. Parker, K.J., Maestripieri, D.: Identifying key features of early stressful experiences that produce stress vulnerability and resilience in primates. Neurosci. Biobehav. Rev. 35(7), 1466–1483 (2011)
7. Norris, F.H., Tracy, M., Galea, S.: Looking for resilience: understanding the longitudinal trajectories of responses to stress. Soc. Sci. Med. 68(12), 2190–2198 (2009)
8. Peres, J.F., et al.: Police officers under attack: resilience implications of an fMRI study. J. Psychiatr. Res. 45(6), 727–734 (2011)
9. Block, J., Kremen, A.M.: IQ and ego-resiliency: conceptual and empirical connections and separateness. J. Pers. Soc. Psychol. 70(2), 349–361 (1996)
10. Connor, K.M., Davidson, J.R.: Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 18(2), 76–82 (2003)
11. Wagnild, G.M., Young, H.M.: Development and psychometric evaluation of the Resilience Scale. J. Nurs. Meas. 1(2), 165–178 (1993)
12. Kruegar, J.: Enhancement bias in descriptions of self and others. Pers. Soc. Psychol. Bull. 24, 505–516 (1998)
13. Lieberman, H.R., et al.: Severe decrements in cognition function and mood induced by sleep loss, heat, dehydration, and undernutrition during simulated combat. Biol. Psychiatry 57(4), 422–429 (2005)
14. McHugh, R.K., et al.: Cortisol, stress, and attentional bias toward threat. Anxiety Stress Coping 23(5), 529–545 (2010)
15. van Wingen, G.A., et al.: Persistent and reversible consequences of combat stress on the mesofrontal circuit and cognition. Proc. Natl. Acad. Sci. USA 109(38), 15508–15513 (2012)
16. Flinn, M.V., et al.: Evolutionary functions of early social modulation of hypothalamic-pituitary-adrenal axis development in humans. Neurosci. Biobehav. Rev. 35(7), 1611–1629 (2011)

17. McGirr, A., et al.: Dysregulation of the sympathetic nervous system, hypothalamic-pituitary-adrenal axis and executive function in individuals at risk for suicide. J. Psychiatry. Neurosci. 35(6), 399–408 (2010)
18. Maier, S.F., Watkins, L.R.: Role of the medial prefrontal cortex in coping and resilience. Brain Res. 1355, 52–60 (2010)
19. Conti, A.A., Conti, A.: Frailty and resilience from physics to medicine. Med. Hypotheses. 74(6), 1090 (2010)
20. Gumbsch, P.: Materials science. Modeling strain hardening the hard way 301(5641), 1857–1858 (2003)
21. DiCorcia, J.A., Tronick, E.: Quotidian resilience: exploring mechanisms that drive resilience from a perspective of everyday stress and coping. Neurosci. Biobehav. Rev. 35(7), 1593–1602 (2011)
22. Russo, S.J., et al.: Neurobiology of resilience. Nat. Neurosci. 15(11), 1475–1484 (2012)
23. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Eur. Heart J. 17(3), 354–381 (1996)
24. Kirschbaum, C., Pirke, K.M., Hellhammer, D.H.: The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology 28(1-2), 76–81 (1993)
25. von Dawans, B., Kirschbaum, C., Heinrichs, M.: The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. Psychoneuroendocrinology 36(4), 514–522 (2011)
26. Dickerson, S.S., Kemeny, M.E.: Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. Psychol. Bull. 130(3), 355–391 (2004)
27. Delahaij, R., et al.: Predicting performance under acute stress: the role of individual characteristics. Int. J. Stress Manag. 18(1), 49–66 (2011)
28. Biondi, M., Picardi, A.: Psychological stress and neuroendocrine function in humans: the last two decades of research. Psychother. Psychosom. 68(3), 114–150 (1999)
29. Restituto, P., et al.: Advantage of salivary cortisol measurements in the diagnosis of glucocorticoid related disorders. Clin. Biochem. 41(9), 688–692 (2008)
30. Burghy, C.A., et al.: Developmental pathways to amygdala-prefrontal function and internalizing symptoms in adolescence. Nat. Neurosci. 15(12), 1736–1741 (2012)
31. Kent, M., et al.: A resilience-oriented treatment for posttraumatic stress disorder: results of a preliminary randomized clinical trial. J. Trauma. Stress 24(5), 591–595 (2011)
32. Courtet, P., et al.: The neuroscience of suicidal behaviors: what can we expect from endophenotype strategies? Transl Psychiatry. Transl. Psychiatry 1 (2011)
33. Combat Stress. U.S. Marine Corps (2000)
34. Ben-Zur, H., Yagil, D., Oz, D.: Coping strategies and leadership in the adaptation to social change: the Israeli kibbutz. Anxiety Stress Coping 18(2), 87–103 (2005)

Part VI

Applications of Augmented Cognition

So Fun It Hurts – Gamifying an Engineering Course

Gabriel Barata, Sandra Gama, Joaquim Jorge, and Daniel Gonçalves

Instituto Superior Técnico / INESC-ID, Rua Alves Redol, 9, 100-029, Lisboa, Portugal
gabriel.barata@gist.utl.pt, sandra.gam@gmail.com,
jorgej@inesc.pt, daniel.goncalves@inesc-id.pt

Abstract. Good games are good motivators by nature, as they make players feel rewarded and fulfilled, which pushes them forward to persist and resist frustration. Gamification is a novel technique that uses game elements like points and badges, to motivated and engage users into embracing new behaviors, such as improving one's health condition, finances or productivity. In this paper, we present an experiment in which an MSc college course was gamified to improve student interest and engagement. The gamified course led to better learning results and participation. However, there were several negative side effects that detracted from the overall experience. We will describe them, identifying their causes and describe possible alternatives to better tailor the gamified experience, stemming from the analysis of the data gathered so far.

Keywords: Education Gamification, Perils, Student engagement, Motivation.

1 Introduction

The use of games in non-game contexts is gaining notoriety during the last years. Known as Gamification, it consists in using game elements, instead of full-fledged games, in non-game contexts [1]. It is typically used to keep users engaged and motivated to adopt and perform specific behaviors [2] which makes it of special interest for marketing [3]. Gamification has also been used for a large variety of purposes, like helping people to eat better [4] or to be more productive [5] or eco-friendly [6]

Gamification emerged as a powerful behavior driver, by exploring the motivational power of games and applying it to other domains. Games make players feel rewarded, fulfilled and satisfied, by making them experience what may be called of flow [7], [8]. Flow is what makes players persist and endure, which explains why World of Warcraft players reported to spend 21 hours per week playing the game [9].

Games have been used as motivators with success in education. In different experiments, students from different academic levels were subject to learning with video games, and significant improvements in subject understanding, diligence and motivation were observed [10], [11], [12]. Good games are natural learning machines [13]. Unlike traditional educational materials, games can deliver information on demand and within context, and are balanced so that players do not become either bored or frustrated. This suggests that games and gamification have a great potential to mold human behavior and help people learn new skills, which is also supported by recent

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 639–648, 2013.

research. Typical gamified applications rely on game elements such as Points, Badges and Leaderboards as the core of the experience, the so called PBL [19]. While leaderboards allow users to compare themselves with others, points and badges are external rewards for completing certain actions. However, relying solely on these external motivators without considering important human factors like the need to feel competence, autonomy and relatedness [20], will not only fail to engage users, but will also overcrowd any existing interest and internal motivation to perform the behavior in hand [21]. Gamification should be used to boost the user's internal motivation [22].

Jigsaw [14], for example is a gamified application that helps users learn Photoshop, through a jigsaw puzzle that challenges players to match a target image. Although no empirical evaluation was presented, users reported being able to explore the tool and discover new techniques. GamiCAD [15] in turn, is a gamified tutorial system for AutoCAD. By performing line and trimming tasks, users help NASA build a spacecraft to participate in an Apollo mission. Tasks are designed to be challenging and users are encouraged to repeat them until they achieve the required score. When compared to a non-gamified version, results show that users completed tasks faster in GamiCAD and found the experience to be more engaging. Lee Sheldon describes [16] how a conventional learning experience can be designed as a game, without using technology, to engage students and make classes more fun and interesting. Students start with an F and go all the way up to an A+, by completing quests and challenges, which will reward them with experience points. Khan Academy [17] on the other hand, is a free online service that allows users to learn about several topics, such as algebra, economics or history, by watching videos and then completing exercises. Their progress is rewarded with energy points and badges. Similarly, Codeacademy [18] teaches online students to code in numerous programming languages, also using points and badges to track their progress.

Gamified examples like these suggest a synergic effect between gamification and education. However, little attention has been paid to how these approaches can negatively influence the students' engagement to learn. In this paper we present an experiment in which a college course, Multimedia Content Production (MPC), was gamified, and the problems we found, pointing to possible solutions. We start by describing the course and both the gamified and non-gamified instances, which were deployed in different academic years. Following will be a discussion of the main effects of gamification over student participation and diligence, and we also address in detail the negative side-effects of using a gamified course. We finish by suggesting a few design guidelines for gamified learning experiences.

2 The MCP Course

Multimedia Content Production (MCP) is a 5-month long MSc course, in the Information Systems and Computer Engineering degree at Instituto Superior Técnico (University of Lisbon). In the non-gamified year, course evaluation comprised five theoretical quizzes (25% of total grade), a multimedia presentation (20%), lab classes (15%), a final exam (35%), online participation on the course's forums (5%) and class

attendance (5% bonus grade). The final grade was a value between 0 and 20. In the gamified instance, a new grading system was introduced, where students participated in a game-like experience and were awarded experience points (XP). The evaluation consisted of quizzes (10%), a multimedia presentation (20%), lab classes (15%), a final exam (35%) and a set of collectible achievements (20%, plus a 5% grade bonus). Most achievements were multi-level, for a total of 75 badges that could be won (as well as the corresponding XPs). Compared to the first year, the evaluation method was similar, with achievements replacing online participation and attendance.

Fig. 1. The MCP Leaderboard

The MCP course was gamified using 6 core game elements: XP, levels, leaderboards, challenges, badges and a skill tree. The leaderboard was the entry point to the whole gamified experience (see Fig. 1). It allowed users to track their progress, explore their own and others' achievement history, and to compare themselves with other classmates. XP and levels served the main purpose of transmitting direct feedback and progress. Students were awarded with XP as they completed course tasks. Every 1200 XP corresponded to a new progress level, in a 20-level scale, which reflected the student's current grade. For example, a student with 12000 XP (10 × 1200) would be at level 10, which corresponds to 10 grade points.

We are aware that some of our game elements do match the PBL formula and work as external rewards. However, we tried to align the goals of the gamified experience with those of the students, which should motivate them by identification and integrated regulation. As posed by the self-determination theory (Deci and Ryan, 2004), these are the most autonomous forms of extrinsic motivation and share some features with the intrinsic forms. We tried to improve three innate needs of intrinsic motivation: competence, by providing positive feedback and displaying progress with points, levels and badges; autonomy, by offering different options of what challenges to pursue and level up; and relatedness, by allowing students feel part of a community and

participate in the forums. We tried to further improve autonomy with the skill tree (where different paths could be followed, and relatedness, by adding challenges to encourage students to cooperate.

3 Playing the MCP Game

Overall, the students did well. From a total of 52 students, six reached level 20 (the maximum possible grade!), with no student below level 14, except for an exchange student, a late arrival that was unable to adapt to the course and school (reached level 9, thus failing the course) and a student that gave up at the middle of the semester. These two students will be excluded from the subsequent analysis. Figure 2 summarizes the experience levels reached by the students, and shows the grades to have improved thanks to gamification, when compared to the non-gamified version of the course.

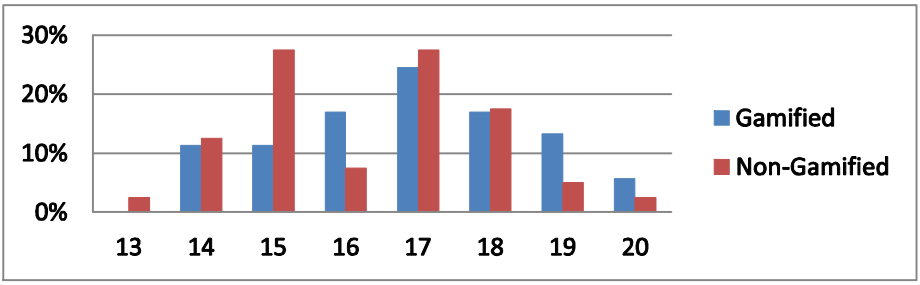


Fig. 2. Percentage of students per final grade

More important than grades, gamification led the students to participate more and be more active learners. Throughout the semester, a total of 2235 posts were made by students, for an average of 139 per week while classes lasted. This contrasts with a much lower figure for the un-gamified version, where only 211 posts *overall* were made by students. As posts were done mostly to gain certain achievements, for which some work was required, this also means that students worked more often on tasks that exercised the skills learned in the course, with a consequent increase in reinforcement learning, made evident in the final grades.

There were, however, big asymmetries between students. Indeed, the relatively high grades were reached in many different ways, sometimes, as we will see, reluctanty! By carefully studying the ways in which different students played the game throughout the semester, we were able to identify the following typical profiles.

— *The Achiever.* Achievers (11 students, 21%) constantly fought for the first place in the leaderboard. Seldom did their position fall below 10th place. These were the students that really enjoyed playing the game, going beyond the minimal requirements just to exercise their skills and have fun.

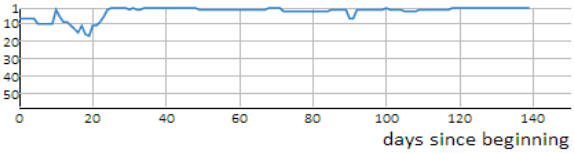


Fig. 3. Typical leaderboard evolution for an *Achiever*

– *The Late Awakener.* Late Awakeners (8 students, 15%) didn't, at first, understand how the course worked. Accustomed to traditional courses with well-defined evaluation moments (a project, an exam, etc.), they neglected the course achievements at first. Once the game progressed and they saw themselves falling behind on the leaderboard, they started participating, often with good results.

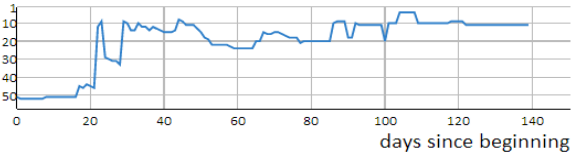


Fig. 4. Typical leaderboard evolution for a *Late Awakener*

– *The Consistent Student.* Consistent students remained roughly in the same position throughout the semester, in the middle-bottom part of the leaderboard. There might be some highs and lows, but they clearly spend a consistent (and not very high) effort with the course. This was the most frequently found category, with 21 students (40%). They typically only went after achievements that were explicitly mentioned in class, with deadlines and, thus, similar to what they know form traditional courses.

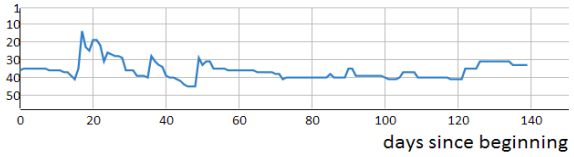


Fig. 5. Typical leaderboard evolution for a *Consistent Student*

– *The Disheartened Student.* These (11 students, 21%) were students that started a strongly at the beginning of the semester but that, after three or four weeks, reverted to a *Consistent Student* behavior of doing the bare minimum tasks explicitly mentioned by the professors.

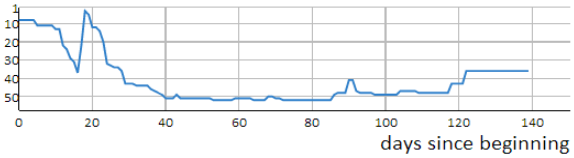


Fig. 6. Typical leaderboard evolution for a *Disheartened Student*

In Fig. 7, we can see how students of different types were spread throughout the leaderboard. It is evident that Achievers and Late Awakeners were the best students, while Consistent and Disheartened appear close to the bottom of the list.

Fig. 7. Final leaderboard position of different student types, (from 1, left, to 52, right)

Throughout the course, the differences between these user profiles were made apparent by the nature of comments by the students and the way they participated. Achievers were clearly driving the game forward very actively. Consistent students, while participating, contributed less to the discussion beyond the posts that would strictly earn them achievements. Even so, we can see (Figure 8) that students of all profiles participated. It must be noted that Achievers were atypical in this regard, participating much more than the others. In fact, a set of t-tests shows statistically significant differences only between Achievers and other profiles, but not between the others (with 95% confidence). This asymmetry led to problems, as we will see below.

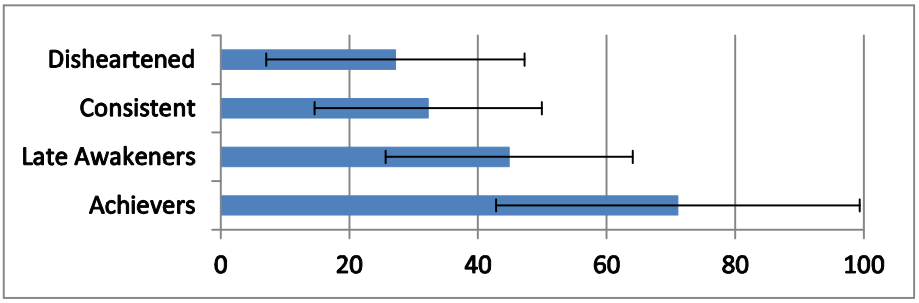


Fig. 8. Average number of posts per student, for the different profiles (error bars: st.dev.)

3.1 Problems with the Game

At the end of the game, we asked students to fill in a questionnaire inquiring them on different facets of the game. We had 45 respondents, out of the 52 students. All questions were based on five point Likert scales.

Students were asked whether they had liked the gamified course. Most rated it positively, as seen in Table 1 (avg=3.51, stdev=1.04). Achievers gave it a higher rating (avg=3.89). Late Awakeners, Consistent and Disheartened students gave it ratings of 3.5, 3.17 and 3.67, respectively. Surprisingly, the students that gave worse ratings to the course were Consistent students. Their ratings ranged from 1 to 5, with five rating it 1 or 2, and eight rating it 4 or 5. This hints at some hidden structure inside this group, not revealed by their leaderboard behavior. Also, it shows that while the Disheartened students appeared to have lost interest in the course, they still liked it more than Consistent students (only one rated it 1 or 2, and five rated it 4 or 5).

Table 1. Questionnaire responses (on a 5-point Likert scale)

		Total	*Achiever*	*Late-Awakener*	*Consistent*	*Disheartened*
Course Rating	**Avg**	**3.51**	3.89	3.50	3.17	3.67
	Stdev	**1.04**	1.27	0.53	1.04	1.00
Motivation	**Avg**	**3.76**	4.22	3.50	3.56	3.78
	Stdev	**1.05**	0.97	0.93	1.15	0.97
Workload	**Avg**	**4.09**	4.89	3.88	3.89	3.89
	Stdev	**0.92**	0.33	0.83	1.08	0.78

A question about how motivating they found the gamified course yielded a similar pattern. With a total average of 3.76 (stdev=1.05), Achievers were by far the most motivated (avg=4.22), followed by Disheartened students (3.78). Future analysis will focus on why students that apparently "gave up" on the course actually liked it more and were more motivated than those that persisted. The questionnaire also had a set of qualitative questions that highlighted the problems discussed below.

Workload. Many users mentioned a high workload as a detrimental factor. When asked to compare the workload in this course to that of others (from 1-much less to 5-much more), they replied with a 4.09 average (Table 1). Achievers rated it higher (avg=4.89!), consistently with their struggle for the topmost places in the leaderboard. However, they seldom complained in the qualitative questions. This can mean they were working more for the pleasure of participating and peer recognition. Still, this was an issue for most students. We were convinced that the amount of work hours needed for this course was not dissimilar to the demands of traditional courses (with large programming projects and other tasks throughout the semester). To address this matter, we asked users about it in a post-questionnaire follow up. Responses varied, but a pattern emerged: it is not only the actual workload but the *perception of workload* that matters. Many courses only require work from students at very limited times throughout the semester (close to a project deadline, an exam, etc.). The gamified course requires them to do much smaller tasks, but requires them continuously. This created the perception that they were "always working for this course", even when the total effort spent was similar to that of other courses.

Comparison Pains. Several students complained about lack of privacy or the visibility of their leaderboard position. They did not want to be compared with, better placed students. Achiever students participated more than could be asked of a typical student. Seeing such a level of activity discouraged others, who felt they could not compete at that level. They resented the fact even while (or, probably, because) getting a better position depended solely on their work. This was exacerbated by the "Talkative" achievement that rewarded classroom participation. Those that didn't participate resented the XP awarded to those that did. Five of the eight students that complained about "Talkative" in the questionnaire were Disheartened students. This reinforces the idea they want to participate, but are intimidated by a level of activity they feel is

beyond their reach. In traditional courses not all students have the same interest and produce same quality work. This, however, happens silently throughout the semester. The gamified course makes it apparent in real-time.

Reward quality, not quantity. The way the game was set up, students were rewarded for the sole act of participating in the several tasks and challenges posted to them. There were no distinctions in terms of the quality of the work produced. They (rightly) felt it was unfair for contributions of different quality to be rewarded similarly.

Awaking too late. Many of the Consistent students only realized they were getting behind once the course was too far into the semester. As many challenges were time-based, it was now too late for them to fully recover, and many didn't try. Looking at individual achievements, they thought that, since each, individually, isn't worth much, there was no point in working for them. Of course, once their colleagues had amassed sufficient XP points making it apparent the achievement XP add up to a significant amount, they wanted to make up for lost time. By then it was too late. This is where the gaming metaphor breaks down: in a computer game, it is possible to reload and try again. In gamified education (and real-life, in general) that is only possible within very limited boundaries. A subdivision of the Consistent group separating "too late awakeners" is probably relevant and will be considered in future analysis.

Competition vs. Cooperation. Many students complained about the course to be too competitive. However, they did not take advantage of the collaborative features in the game. For instance, an achievement rewarded all students in a lab class if they all did well. It was supposed to serve as an incentive to students helping others. In practice, this *never* happened. Instead, students with good lab performance complained about groups with lower performance, as it being "*their fault"* the extra XP hadn't been awarded. This, and similar occurrences, leads us to conclude that, despite the fact they complained about the course being competitive, they are, by nature, competitive, that is, in fact, the culture in our school. Again this was a matter of perception: gamification made explicit that not all students have the same skills (making them resent competition).

3.2 Design Implications

From the problems above derives a set of design implications that should be taken into consideration when gamifying this type of course:

- **Lighten the pace.** The perceived workload must be carefully managed. The intervals between tasks should be carefully chosen to better balance this facet of the game.
- **Careful comparisons.** Consider other leaderboard types that don't make the direct comparison between students of widely different ratings so easy (displaying only the immediate neighbors, having leaderboards for different "leagues", etc).

- **Reward quality.** Estimate the quality of each student's participation and award XPs accordingly. This will increase the amount of work done by the professors but is a requirement for the perceived fairness of the course.
- **Make them participate as soon as possible.** Many students only want to start playing when it is too late. Tailoring the game experience so that they are compelled to participate (and see meaningful rewards) early on will yield better results.
- **Give them the chance to make up for lost time. W**hile some tasks and challenges will always be time-bound, whenever possible it should be allowed for students to address the different challenges in a more unconstrained way.
- **Provide means for cooperation.** These should not be completely decoupled from competition. Find mechanisms where several students can work together towards a common goal but maintain the ability for students can show off their work.
- **Make it all about the game.** Several students thought they could neglect the game as some traditional evaluation components (ex: exam) were still in place. Reducing their importance (or getting rid of them altogether) will dispel this illusion.

4 Conclusions

Education gamification is a growing trend, with clear advantages in terms of student motivation. However, the gamified experience needs to be carefully tailored not only in absolute terms, but also taking into account the culture and specificities of the students and school. We've shown how problems can arise that detract from the learning process. Most problems mentioned above have to do with the timing for the different game elements and related tasks. These have to be carefully adjusted in order to provide a more balanced gaming experience. Next semester, we will deploy a new version of the gamified course, adjusted based on the lessons learned here. We will explicitely measure engagement and characterize the students trying to fine-tune the profiles defined above. We will also assess the influence of each game element.

Acknowledgements. This work was supported by the Portuguese Foundation for Science and Technology (FCT): individual grant SFRH/BD/72735/2010; project PAELife AAL/0014/2009; and project PEst-OE/EEI/LA0021/2011.

References

1. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining "gamification". In: Proc. 15th International Academic MindTrek Conf. Envisioning Future Media Environments, pp. 9–15. ACM (2011)
2. Shneiderman, B.: Designing for fun: how can we design user interfaces to be more fun? Interactions 11(5), 48–50 (2004)
3. Zichermann, G., Cunningham, C.: Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps. O'Reilly Media, Inc. (2011)
4. `http://healthmonth.com/` (February 26, 2013)

5. Sheth, S., Bell, J., Kaiser, G.: Halo (highly addictive, socially optimized) software engineering. In: Proceeding of the 1st International Workshop on Games and Software Engineering, GAS, vol. 11, pp. 29–32 (2011)
6. Inbar, O., Tractinsky, N., Tsimhoni, O., Seder, T.: Driving the scoreboard: Motivating eco-driving through in-car gaming. In: Proc. CHI 2011 Workshop Gamification: Using Game Design Elements in Non-Game Contexts, ACM (2011)
7. Csikszentmihalyi, M.: Flow: The psychology of optimal experience. Harper Perennial (1991)
8. Nakamura, J., Csikszentmihalyi, M.: Flow theory and research. In: Oxford Handbook of Positive Psychology, pp. 195–206 (2009)
9. Yee, N.: The daedalus gateway (2005), http://www.nickyee.com/daedalus/archives/001365.php
10. Coller, B., Shernoff, D.: Video game-based education in mechanical engineering: A look at student engagement. Int. Journal of Engineering Ed. 25(2), 308–317 (2009)
11. Kebritchi, M., Hirumi, A., Bai, H.: The effects of modern math computer games on learners' math achievement and math course motivation in a public high school setting. British Journal of Educational Technology 38(2), 49–259 (2008)
12. Squire, K., Barnett, M., Grant, J.M., Higginbotham, T.: Electromagnetism supercharged!: learning physics with digital simulation games. In: Proc. ICLS 2004, International Society of the Learning Sciences, pp. 513–520 (2004)
13. Gee, J.P.: What video games have to teach us about learning and literacy. Comput. Entertain. 1(1), 20 (2003)
14. Dong, T., Dontcheva, M., Joseph, D., Karahalios, K., Newman, M., Ackerman, M.: Discovery-based games for learning software. In: Proc. CHI 2012, pp. 2083–2086. ACM, New York (2012)
15. Li, W., Grossman, T., Fitzmaurice, G.: Gamicad: a gamified tutorial system for first time autocad users. In: Proc. 25th Annual ACM Symposium on User Interface Software and Technology, UIST 2012, pp. 103–112. ACM, New York (2012)
16. Sheldon, L.: The Multiplayer Classroom: Designing Coursework as a Game. Course Technology PTR (2011)
17. https://www.khanacademy.org/ (February 26, 2013)
18. http://www.codecademy.com/ (February 26, 2013)
19. Werbach, K., Hunter, D.: For the Win: How Game Thinking Can Revolutionize Your Business. Wharton Digital Press (2012)
20. Deci, E., Ryan, R.: Handbook of self-determination research. University of Rochester Press (2004)
21. Rigby, S., Ryan, R.: Glued to games: How video games draw us in and hold us spellbound. Praeger (2011)
22. Deterding, S.: Gamification: designing for motivation. Interactions 19(4), 14–17 (2012)

A Practical Mobile Dry EEG System for Human Computer Interfaces

Yu M. Chi[1], Yijun Wang[2], Yu-Te Wang[2], Tzyy-Ping Jung[2], Trevor Kerth[1], and Yuchen Cao[1]

[1] Cognionics, Inc. San Diego CA 92121, USA
[2] University of California, San Diego La Jolla CA 92093, USA

Abstract. A complete mobile electroencephalogram (EEG) system based on a novel, flexible dry electrode is presented. The wireless device features 32-channels in a soft, adjustable headset. Integrated electronics enable high resolution (24-bit, 250 samples/sec) acquisition electronics and can acquire operate for more than four hours on a single AAA battery. The system weighs only 140 g and is specifically optimized for ease of use. After training users can self-don the headset in around three minutes. Test data on multiple subjects with simultaneously acquired EEGs from a traditional wet, wired system show a very high degree of signal correlation in AEP and P300 tasks.

1 Introduction

Portable electroencephalogram (EEG) based systems have long been explored as a tool for implementing brain- computer interfaces (BCI) [1,2,3,4]. Despite the many advancements in signal processing and algorithms towards realizing a useful system, the EEG headset itself has remained a critical barrier against a practical device. Conventional EEG systems are cumbersome, requiring extensive subject preparation. Recently, dry electrode EEG systems have been explored as an alternative. However, dry headsets still suffer from numerous issues relating to comfort (e.g., hard metal pins) and signal quality. This paper aims to present a new, wireless dry EEG headset that specifically addresses the need for a complete, mobile system and will cover both the design and experimental validation.

2 Sensor Design

Mobile EEG systems have focused heavily on the use of dry electrodes with mixed results. In principle, dry electrodes are attractive due to the lack of scalp preparation. In practice, they have multiple issues relating to signal quality, usability and comfort. Current dry electrodes mostly utilize the straight metal spring-pins structure [5] to push through the hair. Pin based designs introduce significant discomfort and in military or ambulatory applications, pose an injury hazard [4]. Spring loaded sensors are also too intricate and complex to produce

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 649–655, 2013.

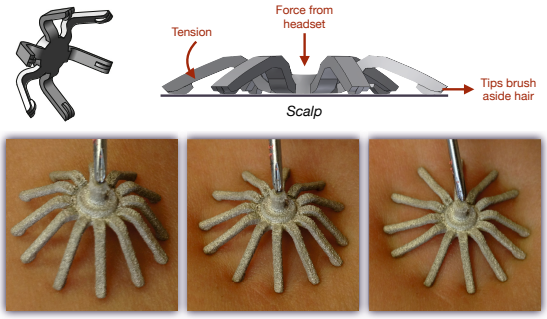


Fig. 1. Cognionics patent-pending flexible dry electrodes. The design consists of angled legs that can deform under pressure enabling penetration of hair without discomfort or risk of injury to the scalp.

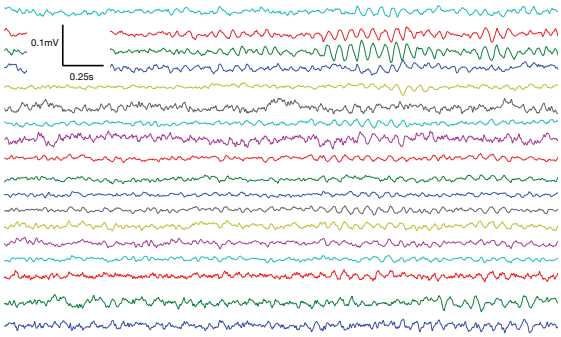


Fig. 2. An 18 channel raw data segment collected by the Cognionics headset using the flexible dry electrodes. The top traces clearly show alpha burst activity and demonstrates the high signal quality of the flexible dry electrode.

inexpensively. Other dry electrode designs exist, primarily based on conductive fabrics [6] or conductive brushes [7]. However, such approaches do not readily penetrate all types of hair and have issues with cost and longevity. Finally, many dry electrode systems also require significant fiddling of both the sensor and cap to generate sufficient pressure, eliminating many of their convenience advantages.

To address the performance and form-factor limitations with conventional dry EEG electrodes (e.g., hard metal pins), Cognionics, has developed a patent-pending, flexible dry electrode (Fig. 1) specifically designed to easily penetrate layers of hair while remaining safe, even under hard pressure. The new dry electrodes utilize a set of angled legs rather than straight pins. The electrode is made from a nylon material (3-D printed) that permits the legs to bend and flex outward under pressure. The flexing action helps push aside strands of hair for better scalp contact with minimal adjustment. Under hard pressure, the entire structure simply deforms and flattens to remain safe. For conductivity, the sensors are coated with metallized paint.

Fig. 3. Cognionics 32-channel dry EEG headset. The headset is made from soft fabric and completely encloses the wiring for the headset. The miniaturized electronics (box at the back of the head) operates from a single AAA battery and contains onboard amplification, digitization and wireless telemetry. Total system weight is only 140 g.

Under normal usage, the legs are only slightly deformed to provide for a minimal tension to ensure adequate pressure on the scalp and should not introduce any discomfort to the user. For users with different hair thicknesses, we have different sized sensors (e.g., broad legs for near-bald, thin for thick hair) to optimize hair penetration and comfort. Under most haired subjects, minimal adjustment is required to achieve sufficient scalp contact and a simple pressing motion is sufficient to part any trapped hair. The metallized legs provide for a sufficiently low impedance contact (100−500 kΩ) to ensure low-noise EEG acquisition as shown in Figure 2.

3 Headset Design

We have designed a soft fabric based head harness (Fig. 3) that is adjustable to a wide variety of heads shapes and can meet the specific requirements of dry EEG systems. The headset consists of self-adjustable straps to easily conform to a variety of head sizes. After training, donning can be accomplished in less than 3 minutes without assistance. The headset is completely self-contained and contains all the necessary electronics and streams data wirelessly via Bluetooth. A very high data quality, comparable to research-grade bench systems, is made possible by 32 simultaneous 24-bit A/D converters with active electrode buffers on each channel. Typical battery life is around 5 hours of continuous streaming using a single AAA battery.

4 Wireless Data Acquisition Electronics

The latest advancements make it possible to construct a very high-quality portable EEG devices [4] that is far smaller than the traditional 'shelf' type systems. The electronics box for the 32-channel headset measures 2.5" x 2.5" x 0.75" and houses

the amplifiers, digitizers, micro controller and wireless transceiver along with a single AAA battery for power. A summary of the system's specification is listed in Table. 1.

Evoked responses have been a mainstay for EEG-based brain computer interfaces. With wireless systems, especially ones based on conventional protocols (e.g., Bluetooth) optimized for reliable data transfer, issues with latency and jitter often prevent the accurate alignment of stimuli, event markers and EEG data.

One solution is to simply attach a physical wire to the headset for minimal latency and jitter-free transmission of event markers, as with traditional wired systems. However, such an approach defeats the purpose of a wireless headset. Cognionics has developed a novel wireless method of transmitting EEG trigger signals and event markers based on infrared and RF based custom transceivers (Fig. 4). As will be shown later, this approach permits fully wireless synchronization of EEG with external stimuli with 'wired-equivalent' performance in terms of latency and jitter. The wireless system has the additional benefit of supporting an arbitrary number of receiving headsets, enabling precision timed group experiments that were not previously possible. Finally this approach does not require the use of a custom wireless transceiver for the actual EEG data, retaining compatibility with any generic Bluetooth (or future wireless standard) device.

Table 1. System Specifications

Channels	32 Active plus Reference and Ground
Amplifier Noise	$< 1\mu V_{rms}$, 1-70 Hz
CMRR	>100 dB
ADC Resolution	24 bits
Sample Rate	250 samples/sec
Wireless	Bluetooth v2.1 RFCOMM
Trigger Latency	300 μs
Weight	140 g, fully loaded
Battery	1 AAA NiMH, 5 hours

5 Testing and Validation

Testing and validating the signal from a EEG system has been a difficult endeavor. While bench tests can accurate measure the performance of the system's acquisition electronics (e.g., noise floor, CMRR), the actual performance on an actual subject is difficult to quantify due to the inherent nature of EEG signals, which are generally not repeatable, and the many variations in human head size, shape, and skin condition.

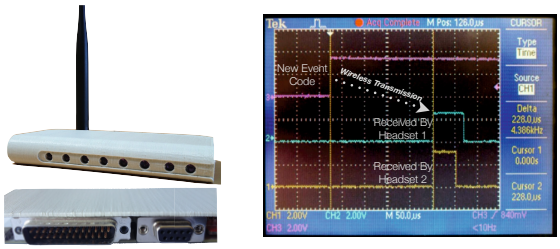


Fig. 4. (left) Cognionics wireless trigger transmitter. (right) Demonstration of novel wireless triggering system. Event codes received by Cognionics trigger unit are received by the wireless EEG headset with a minimal latency (<250 μs) and with virtually no jitter between different systems.

Fig. 5. Photograph of one subject in the test environment. The Cognionics 32-channel dry headset was placed on a subject. Standard wet adhesive electrodes, on top of abraded scalp, were placed adjacent to select dry electrodes and connected to a g.tec amplifier for simultaneous recordings.

An evoked response potential based test protocol is perhaps the best approach since evoked responses offer a repeatable signal. For a fair comparison, a simultaneous recording between wet and dry sensors is needed since conditions may change between recording sessions (e.g., subject fatigue).

For the basic validation experiment, we chose to use an AEP task along with two P300 tasks - one based on static images and another based on video. Since it is impossible to fully cap a subject with both a dry and a wet array, we selected C3-P3 and C4-P4 as the sites of interest for the comparison study.

Figure 5 shows one subject in the test environment. The dry cap was first placed on the subject. Since it is impossible to overlay a wet electrode on top of a dry electrode, the wet electrode must be placed at a location away from the dry electrode with sufficient distance to avoid gel contamination. Physical displacements are not ideal since they change the measured EEG signal. To better simulate a simultaneous wet-dry comparison, two wet electrodes were

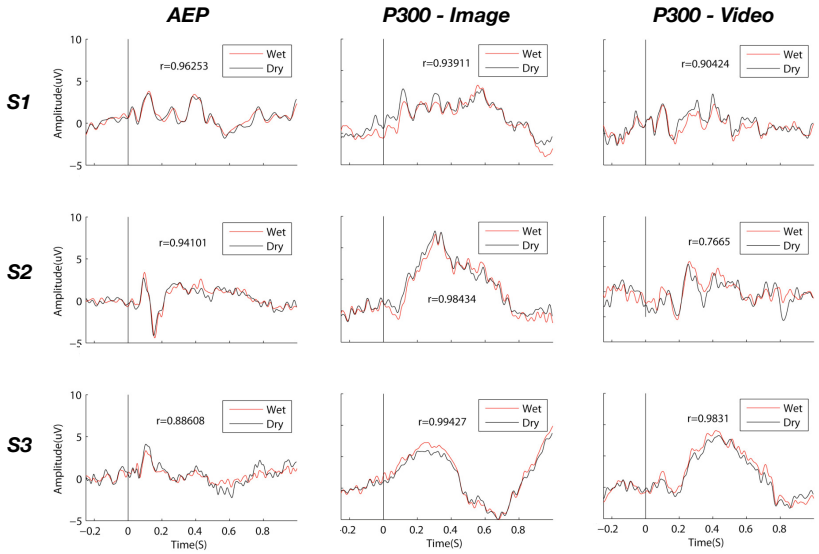


Fig. 6. Exemplary time averaged ERP responses from the three subject in the three tasks (AEP, P300 on still images, P300 on moving video). There is a high correlation and similarity between the signals recorded from the wireless dry system and the standard wet wired system demonstrating the quality of the flexible dry electrode and the accuracy of the wireless triggering system.

placed laterally across each of the dry electrode locations under test (C3, C4, P3, P4). Averaging the two wet electrode 'simulates' a single wet electrode on the exact same spot as the dry electrode for the best comparison. For data acquisition, a g.tec EEG device was used with the wet electrodes.

Three subjects were used for the first validation tests. Figure. 6 shows the time averaged AEP and P300 responses (bipolar C3-P3 montage). Both the wet and dry systems accurately show the expected ERP response. The lack of time shift between the wired wet and wireless dry systems also demonstrate the precision of Cognionics wireless triggering. Almost all of the tests show a high correlation (>0.9) between the wet and dry signals. In the trials with low correlation, the raw signals, as with all of the sets, show a high degree of qualitative similarity.

6 Conclusions

A wireless, 32-channel dry EEG system with novel dry electrodes was demonstrated and tested. The wireless EEG systems includes all of the necessary components for a complete EEG platform, including accurate triggering and event marking. The high quality of the raw signal as well as time-averaged ERP responses demonstrate the viability of the platform for constructing practical mobile brain-computer interfaces.

Acknowledgements. This research was sponsored as part of the DARPA CT2WS program to develop wireless dry EEG headsets.

References

1. Wang, Y.-T., Wang, Y., Jung, T.-P.: A cell-phone-based braincomputer interface for communication in daily life. Journal of Neural Engineering 8(2), 025018 (2011), http://stacks.iop.org/1741-2552/8/i=2/a=025018
2. Oehler, M., Neumann, P., Becker, M., Curio, G., Schilling, M.: Extraction of ssvep signals of a capacitive eeg helmet for human machine interface. In: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2008, pp. 4495–4498 (August 2008)
3. Zander, T.O., Lehne, M., Ihme, K., Jatzev, S., Correia, J., Kothe, C., Picht, B., Nijboer, F.: A dry eeg-system for scientific research and brain-computer interfaces. Frontiers in Neuroscience (2011)
4. Chi, Y.M., Wang, Y.-T., Wang, Y., Maier, C., Jung, T.-P., Cauwenberghs, G.: Dry and Non-contact EEG Sensors for Mobile Brain-Computer Interfaces. IEEE Transactions on Neural Systems and Rehabilitation Engineering 20(2), 228–235 (2012)
5. Matthews, R., McDonald, N., Hervieux, P., Turner, P., Steindorf, M.: A wearable physiological sensor suite for unobtrusive monitoring of physiological and cognitive state. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2007, August 22-26, pp. 5276–5281 (2007)
6. Lin, C.-T., Liao, L.-D., Liu, Y.-H., Wang, I.-J., Lin, B.-S., Chang, J.-Y.: Novel dry polymer foam electrodes for long-term eeg measurement. IEEE Transactions on Biomedical Engineering 58(5), 1200–1207 (2011)
7. Grozea, C., Voinescu, C.D., Fazli, S.: Bristle-sensors–low-cost flexible passive dry EEG electrodes for neurofeedback and BCI applications. Journal of Neural Engineering 8(2) (2011)
8. Chi, Y., Jung, T.-P., Cauwenberghs, G.: Dry-contact and noncontact biopotential electrodes: Methodological review. IEEE Reviews in Biomedical Engineering 3, 106–119 (2010)
9. Ruffini, G., Dunne, S., Farres, E., Cester, I., Watts, P., Ravi, S., Silva, P., Grau, C., Fuentemilla, L., Marco-Pallares, J., Vandecasteele, B.: Enobio dry electrophysiology electrode; first human trial plus wireless electrode system. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2007, August 22-26, pp. 6689–6693 (2007)
10. Sullivan, T., Deiss, S., Jung, T.-P., Cauwenberghs, G.: A brain-machine interface using dry-contact, low-noise eeg sensors. In: IEEE International Symposium on Circuits and Systems, ISCAS 2008, May 18-21, pp. 1986–1989 (2008)

Gamification for Measuring Cyber Security Situational Awareness

Glenn Fink[1], Daniel Best[1], David Manz[1], Viatcheslav Popovsky[2], and Barbara Endicott-Popovsky[3]

[1] Pacific Northwest National Laboratory, Richland, Washington, USA
[2] University of Idaho, Moscow, Idaho, USA
[3] University of Washington, Seattle, Washington, USA
{glenn.fink,daniel.best,david.manz}@pnl.gov,
dr_popovsky@hotmail.com, endicott@uw.edu

Abstract. Cyber defense competitions arising from U.S. service academy exercises, offer a platform for collecting data that can inform research that ranges from characterizing the ideal cyber warrior to describing behaviors during certain challenging cyber defense situations. This knowledge could lead to better preparation of cyber defenders in both military and civilian settings. This paper describes how one regional competition, the PRCCDC, a participant in the national CCDC program, conducted proof of concept experimentation to collect data during the annual competition for later analysis. The intent is to create an ongoing research agenda that expands on this current work and incorporates augmented cognition and gamification methods for measuring cybersecurity situational awareness under the stress of cyber attack.

Keywords: Cyber Defense Competitions, CCDC, cyber defender, cyberwarrior.

1 Introduction

The Pacific Rim Collegiate Cyber Defense Competition (PRCCDC) represents a unique opportunity for observational experiments. While there are many types of observational experiments, in computer security they mostly fall into two classes: laboratory experiments and field studies. Laboratory experiments can be highly controlled and enable researchers to test a hypothesis and quantify the contribution of each of several factors with confidence. With good experimental design, the results may be generalized safely. Unfortunately, the very controls required to obtain certainty cause results to be much less realistic, and potentially less relevant to real life. In contrast, field studies are used in situations where interesting behavior is to be observed, but it is impractical to compare a control group to an experimental group. In field studies, data collected can be highly relevant to real life, but the power of the conclusions that we can draw from these observations is greatly limited because of high variability and contamination from uncontrolled factors. Field studies are typically difficult to replicate, and results may be hard to quantify and merely anecdotal.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 656–665, 2013.

These researchers believe that the PRCCDC, and similar competitions, represent a venue for conducting experiments that are a hybrid of laboratory experiments and field studies. The nature of the competition introduces constraints that (with care) can be adopted as experimental controls while the range of activities available to measure are nearly as unlimited as those that happen in the real world. And possibly just as importantly, the data that can be collected could be published, shared, and reused much more easily without destructive anonymization, unlike that collected in real-world situations. Further, gamification methodologies can be applied that can expand on the purely observational experimentation described in this proof of concept.

2 History of the Collegiate Cyber Defense Competitions

Cyber defense competitions arose out of a military educational requirement for the U.S. service academies [1]. The competition was fierce and the result was so successful that civilian universities began to follow suit. Beginning in 2004, the US Military Academy at West Point adapted their 'capture the flag' exercise to a civilian scenario and introduced the competition at several universities across the country, including the University of Washington which incorporated the event into the Information Assurance and Cybersecurity Certificate program as an annual capstone experience. On February 27 and 28, 2004, a group of educators, students, government and industry representatives gathered in San Antonio, Texas, to discuss the feasibility and desirability of establishing a post-secondary level, national program for cyber security exercises. The outcome of these discussions was 1) a competition architecture with a clear set of rules and roles, 2) a fair and impartial scoring system that provides a level playing field for competitors, 3) an IT infrastructure designed to eliminate possible advantages due to hardware and bandwidth differences at different regional locations, and 4) resolution of possible legal concerns.

The resulting Collegiate Cyber Defense Competition (CCDC) system provides institutions teaching information assurance or computer security a controlled, competitive environment that can assess students' depth of understanding and operational competency in managing and protecting a corporate network [2]. The CCDC helps participating institutions of higher education evaluate their educational programs, provides an educational venue for students to apply the theory and practical skills they learn in their course work, fosters teamwork and ethical behavior, and creates interest and awareness among participating institutions and students. In 2006, the University of Texas at San Antonio agreed to host the first national CCDC. In 2007, the University of Washington opened up their internal competition to outside institutions, establishing the regional PRCCDC as an entrant into the national competition. 2013 is the sixth year of PRCCDC participation in Nationals. There are now ten regional venues: At-Large (virtual) Regional, Mid-Atlantic, Midwest, North Central, Northeast, Pacific Rim, Rocky Mountain, Southeast, Southwest, Western.

During competition, 8-10 student teams comprised of eight students each defend identical networks. The competition lasts 2-3 days. Teams are scored based on ability

to protect and defend against outside threats, maintain availability of web services, respond to business requests, and balance security needs against business needs.

A Red Team of external attackers, often professional penetration testers from local industry, relentlessly attack student networks throughout the competition. Students are expected to resist attack, or recognize and recover from attack, if penetrated. A White Team of judges—in the case of the PRCCDC a team of graduate students from Idaho State University's NIATEC program—issue a series of 'injects,' or administrative chores, that must be accomplished in an orderly and timely fashion in the face of attack. The entire process is designed to simulate the stress and intensity of managing networks in today's hostile Internet environment. These CCDC exercises employ controls designed to preserve fairness and safety among teams from participating schools. These same controls may be used as the foundation of high-quality experimental controls as long as fairness and safety are preserved. For instance, each team begins with a small, pre-configured, operational network they must secure and maintain located on a dedicated internal network. This also allows tight control over competition traffic. Each team is given the same set of business objectives and injects at the same time during the course of the competition.

Each student team is composed mostly of undergraduates, although two at most could be graduate students. No professionals are allowed, and the students may not be currently employed in an IT industry job. Students must be enrolled in a minimum number of class hours to qualify. Faculty advisors are not allowed to be with the team during competition. These restrictions double as experimental controls. The White Team enforces the competition's controls and employs an automated scoring engine that periodically tests availability and function of each student team service and network component during the competition. They also administer and grade responses to injects. Allowing only students and White Team members inside competition rooms eliminates potential variability from the influence of coaches. Running scores are not announced during the competition, eliminating potential stress factors.

The Red Team is the aggressor seeking to disrupt services and business objectives of the student teams. They are non-biased, commercially experienced, and comprised of volunteers. Loose controls are placed on Red Team activities that enforce objectives of fairness and safety. Within these controls, Red Team members employ any attack techniques at their disposal, including non-cyber attacks like social engineering. After the competition, the Red Team usually provides feedback to the student teams on their defenses and how the Red Team attacked them.

3 Data Collection

In this paper, the authors discuss how data that described the effectiveness of collaboration was collected at the PRCCDC. Future studies will include injecting collaboration-enhancing technologies to show the effectiveness of these treatments and augmented cognition methodologies designed to measure participant biological reactions to stress. Data collected was analyzed in a separate publication [3]. In this paper, we discuss experience gained in collecting the data to show the effort required, as well as the benefits this data will be to future studies. An observational experiment

was designed to collect baseline (control) information on collaborative practices in cyber security teams. Collecting full packet traces is common practice at these competitions, but it was felt much more data was needed to tell the stories behind the collaborative interactions that the competition fostered. This section discusses each of the kinds of data collected and how it was collected. During the competition, the following was gathered:

1. Data from the team scoring process,
2. Situational awareness data from team members,
3. Network packets and machine log files,
4. Video and audio of the competition,
5. Stress resilience characteristics of one of the teams.

3.1 Performance Data Capture

Having well defined and fair performance scoring built into the CCDC makes it an excellent source of regular data with a ground truth. Performance and timing data were gathered from the teams' execution of business requirements (injects) that were delivered by email as part of the competition. A HotMail web client was used to record the time when an email instruction was received, opened, and replied to. This timing data was integrated with situational awareness data discussed below. Scoring data gathered included evaluation rubrics for each inject (twenty per team) that guided scoring of student team performance when executing each inject. Computation was done by White Team volunteers and is somewhat subjective. Scoring data was also generated for each successful attack levied against the student teams. Whenever the Red Team infiltrated a student machine successfully, that student team lost points. If the attacked team filed a detailed incident report, they would salvage some portion of their loss. These incident reports helped assess collaborative behavior. Final scores accumulated by each team were gathered from the White Team as an ultimate measure of success. This scoring was partly objective, partly subjective. The subjective part came from humans grading the "goodness" of inject response. The more objective source of data came from the scoring engine which periodically tests the state of all the services teams must maintain. The scoring engine results provided an important source of ground truth when assessing situational awareness.

3.2 Situational Awareness Data Capture

Team situational awareness was measured as a way to infer team performance independently from the competition performance scoring. Researchers, armed with digital audio recorders, were assigned to occasionally ask situational awareness questions of student and Red Team members. Timing and accuracy data were used from their responses and from the injects to conduct an assessment of team situational awareness using Durso's Situation Present Assessment Method (SPAM) [4].

The Questions. The questions used for assessing situational awareness were binary choices (yes/no, A/B) designed to assess the team's cognition of their situation without interrupting their tasks. Reducing interruptions was one of the reasons Durso's model was chosen over interruption-based protocols like Endsley's Situation Awareness Global Assessment Technique (SAGAT) [5]. Additionally, the research team kept questions simple to answer using known, ready-to-hand, materials.

There were seven student teams and one Red Team in the competition. Four researchers gathered data. Each student team was queried every 20 minutes. The Red Team was also queried periodically, but the objective here was to inform the questions of the research team rather than to measure situational awareness. One researcher stayed with the Red Team, the remainder queried student teams. A question matrix was designed for the student teams with one-third of the questions, each, concentrating on concerns of the past 20 minutes, the present, or future 20 minutes, respectively. Durso's work shows that future-oriented questions were most indicative of expertise, so the tense of a question was controlled carefully. The following taxonomic breakdown of question types was used:

1. Defense-related
   a. Policies: What defensive actions should happen?
   b. Priorities: What defensive actions are most important?
   c. Events: What defensive actions were taken?
   d. Causes: What caused or would cause defensive action X?
2. Threat-related
   a. Policies: What offensive actions should happen?
   b. Priorities: From an attacker's view, what is the most important action?
   c. Events: What offensive actions happened or will happen?
   d. Causes: What caused or would cause attackers to take offensive action X?

From this taxonomy, a list of 48 questions was generated. The research team met approximately every 20 minutes and randomly selected one of these questions and applied it to the current situation, filling in information as needed. For example, one question was, "Do you expect your X service to be a likely attack vector in the next 20 minutes?" Before using this question, researchers had to replace X with the name of a service (e.g., email, web, ftp, etc.) thought most fitting at the.It was important to administer the same question to all the teams so as not tip a team off and provide an advantage. For instance, asking whether or not a team had changed the default router password might inform them that they should do this when they had not known to do so on their own.

The Querying Protocol. Each researcher was given the task of querying 2-3 student teams, selected at random, during the remainder of the 20-minute segment. Researchers were instructed to try to approach a team member they had not approached. This induced as much variability into picking the subject as possible. Some teams chose a spokesperson to handle all queries. In that case, the researcher noted the policy and always approached the spokesperson, honoring the team's wishes. The intent is to infer *team* situational awareness from these queries.

To ask a question, a researcher would approach a participant and place a green question card face down on the table in plain view. He then would start his audio recorder and say, "Excuse me, I have a question when you are available." When the participant was ready to answer, he or she would turn over the question card, and the researcher would ask him/her to read the question aloud and answer it. The audio recorder was left running from the initial "excuse me" until either the participant finished answering or five minutes of silence elapsed. At a maximum time of five minutes, the researcher would stop the recorder, pick up the question card, and move on.

The Analysis Plan. Durso's method [4] was employed to measure both situational awareness and workload. In Durso, the time from when the researcher says, "excuse me" until the interviewee reads the question is a measure of workload. Similarly, the time from when the question is read to when the participant answers is a measure of situational awareness. Durso never made any claims about this method working for assessment of team situational awareness. The authors believe effectiveness can be inferred from individual situational awareness, but teams are another matter. There are also difficulties that arise because the venue is not a tightly controlled experiment. Many uncontrolled distractions are happening in a competition whose effects may be larger than the situational awareness effects being measured. This lack of control is inherent to the venue, but the authors believe that quality data can be extracted and generalized, keeping in mind these limitations. Increased statistical power is gained by running the experiment simultaneously on seven teams; however, it must be kept in mind that these are not and cannot be true replications because they are different subjects and are not truly independent.

3.3 Network and Log File Collection

To provide the most information available about network activity, full packet traffic was captured in several key locations. The network topology consisted of a core router connecting all teams to the scoring server and the Red Team (see Figure 1).

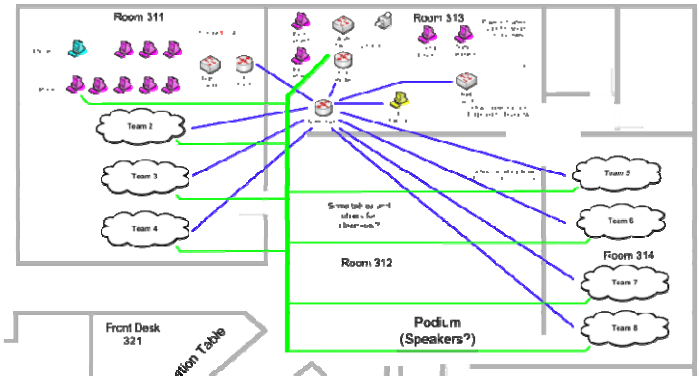


Fig. 1. Competition Network

Connected to the core router, each team's router defined the team's local network. Because Red Team activity could disable a team's router, there was no guarantee that each team's traffic would always reach the core router throughout the event. The aim was to gather as much data from the network, given configuration limitations.

To be as unobtrusive as possible, the core router and team routers were configured to mirror a set of ports to an available port (the "span port"). The associated network interface controller (NIC) of packet-capture laptops connected to the span port were configured to **not** have an IP address—making them essentially invisible. *tcpdump* was configured to capture full packets (headers and all data) by setting the *snaplen (-s)* parameter to 0 (no size limitation). Packet data was output to files of 100 million bytes (-C 100) to ease processing later. A startup script was installed to initialize *tcpdump* and ensure existing packet capture files were not overwritten when the program started. Each machine ran 32-bit Ubuntu Server 9.10 OS, configured with no optional services, in order to minimize attack vectors.

The core router was configured to capture packet data, and because of resource limitations, only three other packet-capture machines were provided on other routers. To allow for possible correlation of network data with captured video, the router of the single team who agreed to be filmed during the competition was one of those. Other packet capture locations were some of the other student teams, the Red Team, and machines teams used to access the Internet for patch downloads. After the event, log data was harvested from all available machines.

3.4 Video and Audio Data Capture

In addition to performance, situational awareness, and network data, video and audio were captured from the competition. City University of Seattle filmed the entire event and provided access to their raw footage. This footage was particularly useful to record the Red team's brief-back at the end of the competition; however, during the body of the competition, the coverage was too uneven as a reliable data source. Not all teams consented to recording which would have been prohibitively expensive in both equipment and time to analyze, so resources were concentrated on the one team from the UW iSchool which graciously agreed to allow video and audio capture.

Eight Logitech 600 webcams were placed strategically within the iSchool team's area to capture interactions and collaboration among participants. The cameras were pointed across the table to capture several subjects at once, allowing a clearer view of team interactions. The team sat in two circular pods with cameras mounted to the table and tops of equipment, facing back across the tables. Camera orientation was periodically checked to make sure they were still aimed correctly.

A single workstation streamed video from all eight webcams using the Logitech camera software and Debut video capture software to capture multiple streams, simultaneously. Eight simultaneous streams of 15fps video were captured at 1280x1024 pixel resolution. While not high quality, this was sufficient to identify whether people were collaborating and a little about their gestures and activities. Since webcams were unable to record clear audio, extra voice recorders were used on each table. During

analysis, a single audio track was used to simplify reviewing the video. To facilitate time synchronization, a sync signal was used to start recording and periodically throughout the competition: a researcher clapped his hands in front of the camera.

3.5 Stress Resilience Characteristics

The student team filmed also consented to being tested, individually, prior to the competition. This was done in order to characterize their psycho-physiological profile as an indicator of their nervous system type. Four tests were given that measured stress resilience, the ability to context-switch, and the ability to maintain balance in their psychological processes under stress. Results led to individualized profiles that, in a business setting, could be useful in managing performance.

This suite of tests was developed by E.P. Ilyin and has proven effective in assessing a subject's ability to handle stress in a variety of occupational settings for particular professions [6,7,8,9,10,11]. Application of this methodology has been helpful in optimizing individual performance in a range of competitive professional environments, including world class sports venues. The authors are adapting this approach to cyber defense competitions. It is believed it could have relevance for developing profiles of effective cyberwarriors, as well as stratagems for identifying and preventing burnout of cyberdefenders stressed by managing networks under constant attack.

3.6 Dry Run

Two dry runs of the data collection technology were conducted to determine feasibility. There were multiple area dependencies where data collection could be derailed. Although some data was lost, the research team was satisfied that a great quantity of useful data was captured. Due to equipment costs and space constraints, the researchers were unable to provide much duplication of collection.

4 Potential Uses of the PRCCDC Data

This data is a "gold mine" of potential research benefits. First, obtaining a realistic set of network data that does not have to be anonymized meets a crying need of the cyber security research community. (In previous research, unavailability of strong anonymization techniques was an important reason why organizations did not share their cyber data and learn from one another's mistakes [12]). Further, research groups at PNNL have long expressed interest in a data set where cyber and video data could be correlated to evaluation of levels of fatigue and stress related to cyber operator error. These authors anticipate using this data to evaluate key characteristics of effective cyber defense teams and individuals. It is expected that the team will return to this data set, again and again, as research matures.

5 Hindrances in Using PRCCDC as a Data Collection Venue

There are some problems discovered in using PRCCDC events as data sources. This is a high stress venue that allows students to impress potential employers and earn a berth to compete at the national CCDC in San Antonio, Texas. Some participants might feel some anxiety knowing that they are being monitored during the competition and not perform optimally.

Since these events are competitions in their own right, not simply experiments, the research team was constrained by the official competition rules. Additionally, the researchers were constrained to ensure that they did not disadvantage, or advantage, any single team by introducing a treatment.

While extremely helpful, those who set up and ran the competition had other jobs and priorities, making it difficult to impose the rigor needed to collect quality data when it impacted people who were not given any incentive to help. Despite these hindrances, the PRCCDC and similar CCDC events remain extremely valuable sources of data.

6 Future Work and Conclusions

This was a pilot study that provided a baseline for future work. The authors plan to interpose collaborative enhancement technology such as Vulcan, designed to improve analyst performance across competing teams, taking care not to (dis)advantage any team. Additionally, different interview techniques and different methods of query delivery and notification are planned to measure the effectiveness of collaboration. Further, semi-structured interviews, or other data sources such as physiological stress measurements, could be introduced to enrich the data set, facilitating the development of a useful profile of an effective cyber warrior.

The contributions of data collection and experimentation with this current work are:

6. Made available a source of de-identified cyber data for publication and sharing.
7. Put forth data-collection practices that may contribute toward a future standard.
8. Identified a new venue for profitable data collection.
9. Contributed towards better quality scientific methods in cyber security research.

These efforts will help researchers for years to come. Benefits of this study are expected to accrue to cyber security workers and researchers into the future.

Acknowledgements. This work was supported by the I4 Initiative of the Pacific Northwest National Laboratory, Richland, WA, managed for the US Department of Energy by Battelle Memorial Institute under Contract DEAC05-76RL01830. The authors wish to thank the organizations who sponsored the PRCCDC, without whom this event would not have been possible: the University of Washington Center for

Information Assurance and Cybersecurity, Idaho State University, Highline and Whatcom Community Colleges, DeVry University, Black Hat, The Boeing Company, Cisco, and Microsoft—in addition to tireless volunteers and student teams from all of the participating schools.

References

1. Schepens, W., James, J.: Architecture of a cyber defense competition. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 4300–4305. IEEE Press, New York (2003)
2. National Collegiate Cyber Defense Competition, http://www.nationalccdc.org/
3. Malviya, A., Fink, G., Sego, L., Endicott-Popovsky, B.: Situational awareness as a measure of performance in cyber security collaborative work. In: IEEE 8th International Conference on Information Technology, pp. 937–942. IEEE Press, New York (2011)
4. Durso, F., Dattel, A.: SPAM: The real-time assessment of SA. In: Banbury, S., Tremblay, S. (eds.) A Cognitive Approach to Situational Awareness, pp. 137–154. Ashgate Publishing, Burlington (2004)
5. Endsley, M.: A methodology for the objective measurement of pilot situational awareness. In: AGARD Symposium on Situational Awareness in Aerospace Operations, pp. 1–9. Neuilly Sur Seine, France ((1989)
6. Il'in, E.P.: Strength of nervous system and methods for this research. In: Psycho-Physiological Fundamentals of Physical Education and Sports, pp. 5–9. Leningrad (1972)
7. Il'in, E.P.: Instant-method for defining the degree of expressiveness of mobility/rigidity of acceleration/deceleration. In: Psycho-Physiological Fundamentals of Physical Education and Sports, pp. 16–21. Leningrad (1972)
8. Dimitrov, A., Popovsky, V.: Influence of several psychological indexes in the performance of fundamental game moves. Coaching Thought, No. 12 Bulgaria (1987)
9. Kuramshin, U., Popovsky, V.: Find your talent. Leningrad, Lenizdat (1987)
10. Kuramshin, U., Popovsky, V.: Prediction of sports abilities in a system of sports orientation for children and adolescents at their residences. Lesgaft University, Leningrad (1985)
11. Popovsky, V., Endicott-Popovsky, B.: Physical culture pedagogical system. In: III International Congress: People, Sport and Health, St. Petersburg, pp. 19–21 (2007)
12. Fink, G., McKinnon, A., Clements, S., Frincke, D.: Tensions in collaborative cyber security and how they affect incident detection and response. In: Seigneur, S., Slagell, A. (eds.) Collaborative Computer Security and Trust Management, pp. 34–63. Hershey, IGI Global (2009)

Human-Robotic Collaborative Intelligent Control for Reaching Performance

Rodolphe J. Gentili[1,2,3,*], Hyuk Oh[1,2], Isabelle M. Shuggi[1,4], Ronald N. Goodman[5], Jeremy C. Rietschel[5], Bradley D. Hatfield[1,2], and James A. Reggia[6,7]

[1] Department of Kinesiology, Cognitive Motor Neuroscience Laboratory, University of Maryland, College Park, MD 20742, USA
[2] Neuroscience and Cognitive Science Program, University of Maryland, College Park, MD 20742, USA
[3] Maryland Robotics Center, University of Maryland, College Park, MD 20742, USA
[4] The Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
[5] Veterans Health Administration, Maryland Exercise and Robotics Center of Excellence, Baltimore, MD 21201, USA
[6] Department of Computer Science, University of Maryland, College Park, MD 20742, USA
[7] Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA
rodolphe@umd.edu

Abstract. In most human-robot interfaces, the user completely controls the robot that operates as a passive tool without adaptation capabilities. However, a synergetic human-robot interface where both agents collaborate could improve the user's performance while reducing the cognitive and physical workload. Specifically, when considering this framework applied to rehabilitation, we examined a shared collaborative control between a human user and an adaptive biologically inspired neurocontroller in order to perform reaching movements with a simulated prosthetic arm. When this neurocontroller was enabled, it progressively learned from the user to control the prosthetic arm, increasing its role in the shared performance and facilitating the user's reaching movements. This resulted in the user's performance enhancement and in a reduction of his/her cognitive workload. The long term goal of this work is to contribute to the development of the next generation of intelligent human-robotic interfaces for rehabilitation.

Keywords: Human-machine/robot collaborative performance, intelligent control, adaptive systems, arm reaching, assistive technology, prosthetic arm, rehabilitation.

1 Introduction

Currently, in most human-machine interface applications, the user fully controls every aspect of the machine performance, which is thus considered as a passive tool

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 666–675, 2013.

controlled in a unidirectional manner with no or very limited capability of adaptation to the user and/or to the environment. However, a more optimal interaction between the user and the machine, such as a robotic limb (e.g., a human controlled robotic arm or finger), would be a dynamic, active and bidirectional process. Therefore, developing a symbiotic human-robot interaction where both the user and the robot can co-adapt and/or cooperate could provide several advantages such as the reduction of ergonomic challenges due to physical and cognitive load, while improving efficiency, quality and safety. Specifically, in the area of rehabilitation, the robotic device that interacts with human can take the form of an assistive device such as a prosthetic limb.

Generally, the working principles of prosthetics as well as many assistive technologies for severely disabled individuals are based on the decoding of available biosignals (e.g., muscle, brain activity, eye, head, tongue movements). These signals are recorded from the user and quantified in order to control the device of interest (e.g., [1-9]). The optimal patient specific interface guides the selection of biosignals that may be employed; e.g., eye, head and tongue movement, and muscle or brain activity [2,5,6,8,9]. Regardless of the interface and the type of biosignals, the user is generally expected to adapt his signal of command in order to unilaterally control the prosthetic while the control system of the device has no or very limited adaptive capabilities [10,11]. While the final aim is to maximize the recovery of motor functions, the available biosignals offer a limited channel of communication to control the prosthesis and/or the assistive device resulting in tedious training, increase of user's fatigue, frustration and cognitive workload as well as a decrement in performance [1,9-12]. It seems reasonable to expect that a prosthetic or an assistive device that would incorporate some adaptive capabilities would reduce the user burden while improving human performance.

Although several investigations proposed adaptive systems to control wheelchairs (e.g., [13,14]), only a few studies have examined biosignals-based intelligent interfaces to control upper limb prosthetics that are critical for the user to perform reaching and grasping task in order to regain interaction with his/her environment. Notably, few previous works have proposed to integrate adaptive elements in the interface to facilitate the decoding process of the control biosignals [10,11]. For instance, Sanchez et al. (2009) employed a reinforcement learning method to adapt the decoding process of invasive brain signal to enhance the control of a robotic arm by a rat [10]. Also, Pilarski et al. (2011) used a similar approach to enhance EMG decoding from human muscles to control a robotic arm [11]. Although very interesting, these previous studies were centered on the decoding process per-se without focusing on the downstream processes related to the controller of the prosthetic device itself. As such, there is a need to develop intelligent collaborative control between the user and a prosthetic arm controller itself. In this regard an adaptive bio-mimetic neurocontroller offers a promising area for developing enhanced human-shared collaborative performance.

Therefore, we propose a human-robotic adaptive collaborative control scheme that provides emergent assistance to the user while performing a reaching task with a virtual prosthetic arm displayed on a computer screen. Using head motion as the

biosignal to control the virtual prosthetic arm, an adaptive biologically inspired neurocontroller will progressively learn to compute the inverse kinematic of the prosthetic limb in order to perform reaching movements towards multiple targets. We predict that the user's performance will be facilitated with concomitant reduction in cognitive workload and frustration as the neurocontroller learns to control the prosthetic arm autonomously. The implications of this approach in the context of intelligent human-robotic interfaces for rehabilitation are discussed.

2 Material and Methods

2.1 The Human-Robotic Interface

The human-machine interface was composed of two elements. The first component acquired the signals from two infrared sensors placed on the head (one on the forehead and one on the chin) of the participants. The movements of the forehead sensor provided the up/down and right/left desired direction from the user whereas the chin sensor was used for selecting/confirming the target acquisition by opening the mouth. Through the movements of these two markers, a motion capture camera-based system (Optotrak™) detected the selection of the target and the desired directional displacement from the user. This information was then used to move a virtual prosthetic arm in a two dimensional workspace displayed on a computer screen that was placed in front of the participant (Fig. 1). It must be noted that as a first step, this study considered a virtual prosthetic arm that was modeled at the kinematic level. However our approach can be employed including an enhanced model of the kinematics and dynamics of the prosthetic arm. In order to ensure consistency, the same targets (same positions, same sequence) were presented to all participants. Once the target was selected by the user, he/she executed (up, down, left or right) head movements that were decoded and provided to the prosthetic arm that moved in the corresponding directions in order to reach the selected target.

The second component of this human-robotic interface included a biologically inspired neurocontroller that functionally reproduces the premotor/motor cortical regions in order to learn an inverse kinematic mapping. In particular, this neurocontroller was able to provide an accurate, robust and efficient inverse kinematics computation reproducing similar kinematics to those observed in human during arm/finger reaching task, while efficiently handling tools, unexpected perturbations, online reacquisition of the targets during simple single reaching motion, as well as more complex movements. These results were obtained with anthropomorphic arms including multiple degrees of freedom as well as with fingers having a mechanical coupling of the last two joints. Although, a simple planar arm with two degrees of freedom arm was considered, this type of neurocontroller can efficiently operate with simulated as well as actual robotic systems such as humanoid arms and fingers that include more complex kinematic mechanisms [15-21].

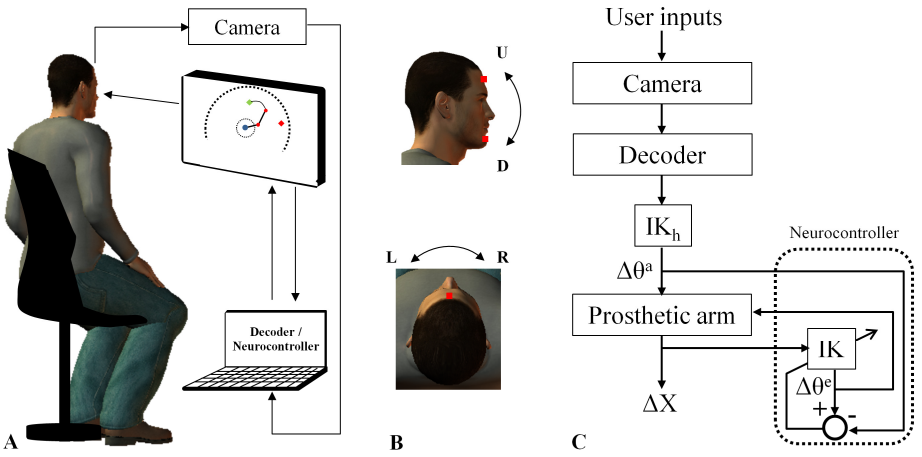


Fig. 1. Working principles of the human-robotic interface. (A). Experimental set-up with the user facing the virtual prosthetic displayed on a computer screen. (B) Marker placed on the forehead to detect the upward (U), downward (D), left (L) and right (R) direction as well as the marker placed on the chin to select the target (C). Human-robotic interaction scheme that allowed adaptive shared control. $\Delta\theta^a$; $\Delta\theta^e$, ΔX represent the actual joint, the estimated joint and the spatial displacement of the prosthetic arm, respectively. IK: Inverse kinematic (h: heuristic).

When considering the present human-robotic framework, the general computational principle of this neurocontroller is to learn an internal representation of the inverse kinematics (i.e., inverse model) of the virtual prosthetic arm by progressively encoding a mapping between its spatial and joint displacements. Thus, when the user moved his/her head, the corresponding (horizontal or vertical) movement directions were decoded and provided to a local inverse kinematics heuristic in order to obtain the corresponding joint displacements and move the virtual prosthetic arm. Simultaneously, the corresponding joints and spatial displacements of the prosthetic arm were provided to the neurocontroller in order to learn the inverse kinematics representation as the user executed reaching movements. As the user moved the prosthetic arm in the workspace, this neurocontroller performed action-perception cycles during which it generated an estimate of the motor commands to move the prosthetic arm in order to reach the targets selected by the user. As the session progressed, the number of movements performed by the user increased and provided further information to the neurocontroller that gradually learned the internal inverse kinematic model of the prosthetic arm by integrating visual (spatial position of the prosthetic arm), and proprioceptive (joint angles of the prosthetic arm) information, as well as internal information related to the neurocontroller. Based on these spatial displacements, the cortical model estimated the joint angles that were compared to the corresponding actual joint movements, providing an error signal that guided the adaptation of the cortical network (for further details see, [15-19]).

2.2 Participants and Reaching Task

Fourteen healthy individuals participated in this study composed of a primary reaching and a secondary cognitive task under various conditions. Only the reaching task that was performed under two conditions will be presented. During the first and second conditions, the subjects had to control, through limited head motion, the prosthetic arm to reach multiple targets while the adaptive neurocontroller was disengaged (i.e., passive prosthetic mode) and engaged (i.e., active prosthetic mode), respectively. Thus, in the first condition (or passive mode), the user exerted traditional control over the prosthetic since he/she fully controlled the prosthetic device that could be considered as a passive tool. During the second condition (active mode), by integrating the user's performance data, the adaptive neurocontroller of the prosthetic arm progressively learned to perform reaching movements towards the targets.

Before starting the experiment and in order to minimize any training or adaptation effects from the user; all the participants went through a familiarization stage where they had to move the virtual prosthetic arm with the neurocontroller disabled and enabled until they felt comfortable in controlling the device. Then, the participants completed two sessions, each of them corresponded to one of the conditions. The condition chosen for the first and second sessions was randomly selected and counterbalanced among the participants. In both sessions, a target (red diamond) to reach was presented on the computer screen within the 2D workspace to the subjects. They had to: i) select/confirm the target acquisition (the target turned green once selected) and then ii) guide the prosthetic arm towards the selected target. Once the participants reached the selected target, the subsequent target was presented and all the information from the previous trial was erased. Each session included 60 trials. To ensure consistency between the two sessions, the sequence of targets to reach was the same during the two sessions (although different from the target set employed during the familiarization phase). The information related to the performance was analyzed throughout each session and for each trial.

In order to assess the quantity of information provided to the prosthetic from the user, the occurrence of head movements were quantified as control signals. Also, the movement time was recorded, the smoothness of the movement path was assessed by means of the jerk [22] and both the linear and angular kinematics of the prosthetic were analyzed. Once each session was completed, participants were requested to complete the NASA TLX questionnaire in order to assess the level of task difficulty and cognitive workload for each task [23]. The indicators of reaching performance (occurrence of head movements, movement time, jerk and the weighted respective role in the performance) were tested using ANOVA. The Huynh-Feldt correction was applied when sphericity was violated [24]. The NASA TLX questionnaire scores were contrasted using paired *t-test* or *Wilcoxon* depending if the assumption of normality was violated or not.

3 Results

3.1 Reaching Performance

Overall, the findings revealed that the user's reaching performance with the prosthetic arm in the passive condition (i.e., neurocontroller disengaged) was inferior to that during the active condition (i.e., neurocontroller engaged).
When comparing the respective roles of the human and of the neurocontroller performance, it appears clearly that the human kept full control of the prosthetics arm in the passive mode and thus produced the entire trajectory (see Fig. 2, upper row). In the active mode, the neurocontroller became progressively dominant in generating the trajectory to reach the targets (see Fig. 2, lower row) and thus gradually reduced the need for user intervention from early to late learning (compare the black and gray portions of the path in Fig. 2). When comparing to the active mode, the passive mode revealed more jerky and irregular movement's paths (Fig.2). Namely, the occurrence of head movements, movement time and jerk values were larger in the passive compared to the active mode ($p<0.001$; Fig. 3A-C).

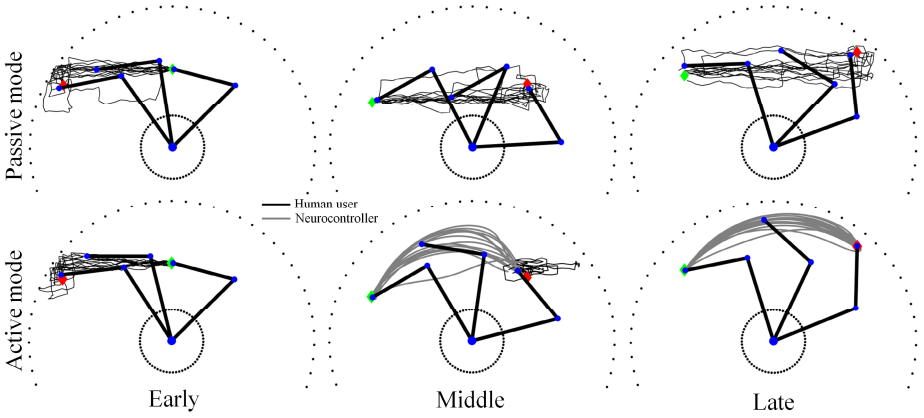


Fig. 2. Reaching performance with the prosthetic arm in the passive (upper row) and active (lower row) mode. The red and green diamonds represent the starting target and the target to reach, respectively. The dotted circular shapes represent the outer and inner limits of the workspace. The black and gray lines represent the portion of the trajectory generated by the human user and by the neurocontroller, respectively.

When focusing on the changes within the session itself, the findings revealed that in the passive mode, the performance was generally stable although towards the end of the session the movement time and smoothness increased and decreased, respectively. The same analysis, conducted in the active mode revealed that the occurrence of head movement required to control the prosthetic arm as well as the movement time were significantly reduced whereas the smoothness of the movement was significantly increased ($p<0.001$; Fig 3A-C). When comparing the respective roles of the human and of the neurocontroller during reaching with the prosthetic arm

in the active mode, the role of neurocontroller, which learned from the user, became progressively preponderant in generating the trajectory to reach the targets. In turn, this resulted in a gradual reduction of the role of the user in controlling the trajectory. Thus towards the end of the session, the user mainly had to control the target selection while the trajectory was generated by the neurocontroller (Fig. 3D). During the passive mode, no change was observed since the users fully control the prosthetic arm at all time.

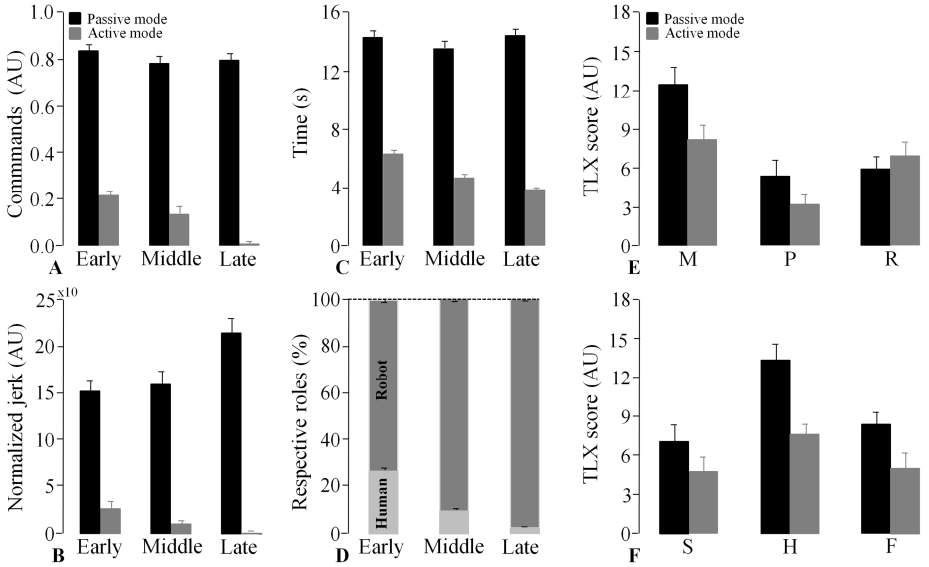


Fig. 3. Indicators of reaching performance along with the cognitive workload and task difficulty assessment during the control of the prosthetic arm in the passive (black color) and active (gray color) mode for the early, middle and late session. (A) Occurrence of head movements, (B) Movement smoothness, (C) Movement time, (D) Respective role in the control of the prosthetic arm during reaching movements, (E-F) NASA TLX scores to assess the mental (M), physical (P) demand, the sensation of being rushed (R), of performing successfully (S), of the task difficulty (i.e., hard or not; H) and the level of frustration (F).

3.2 Cognitive Workload and Task Difficulty

Overall, the NASA TLX results revealed higher scores for the passive compared to the active mode. Specifically, compared to the active mode, the mental demand, the perception to perform successfully, the difficulty to perform the task and the level of frustration were all significantly higher ($p<0.05$). Also, a tendency showed that the physical demand tended to be higher for the passive compared to the active mode ($p=0.06$). The same comparison did not reveal any significant difference between the two modes for the sensation of being rushed ($p>0.73$).

4 Discussion and Conclusion

Overall, the findings suggest that, the cognitive (e.g., mental workload, task difficulty, frustration) and physical effort from the user were reduced whereas the performance was considerably increased (e.g., reduced movement time, increased smoothness) when the neurocontroller was engaged (active mode) compared to the condition where the user fully controlled the prosthetic (passive mode). This finding is in agreement with previous studies that revealed that a collaborative control scheme for wheelchair navigation improved the performance while decreasing the cognitive workload of the user [13, 14].

Specifically, when this adaptive neurocontroller was enabled, throughout the entire session it learned, from the participant, to progressively control the prosthetic arm resulting in an emerging increased assistance to the user to reach the targets. Although the control was shared between the user and the neurocontroller during the entire task, the weights of their respective role evolves as the neurocontroller learned to control the prosthetic arm and thus gradually changed the dynamic of the collaborative effort. Thus, at the beginning of the session, the role of the user in this collaborative framework was predominant since he/she had to control both the target selection and the trajectory of the prosthetic arm. However, as the cortical architecture learned to control the prosthetic device, the roles of the user and of the robot in controlling the trajectory were progressively reversed (i.e., reduced and increased, respectively). Thus towards the end of the session, the user mainly controlled the target selection (i.e., the goal) while the neurocontroller generated the trajectories. In other words, the lower-level aspects of the task, such as the control of the trajectory, were progressively outsourced from the user to the neurocontroller whereas the human user maintained the control of higher-levels aspects of the task such as target selection/movement initiation. Such outsourcing from the human to the robot translated into enhanced performance while the user's cognitive and physical load was reduced. This approach has several implications for users employing prosthetics and assistive devices. First, prosthetics/assistive devices that are based on decoding of biosignals offer a limited communication channel since the recording and interpretation of these biosignals can be complex [1,9]. In addition, the control of such devices generally require long training hours, elevated cognitive workload, and sustained concentration [1,10,12]. By outsourcing some lower-level control features of the task, such as trajectory control, our approach has the potential to develop prosthetic control systems that allow more complex performance while limiting the control of the user to the higher level aspects of the performance (e.g., control related to the goal). This would allow: i) execution of ecologically valid complex movements by the collaborative robot and ii) maintaining a low level of the user's cognitive workload. This is in accordance with previous studies that suggested that the goal control method is a promising option to increase the utility of neuroprosthetics [9, 25, 26]. Second, in daily life, even if the user can correctly control the prosthetic device, this may be at a very high cognitive cost thus reducing cognitive reserve. Under such conditions, the user would not be able to maintain a conversation or deal with unexpected events (e.g., someone inadvertently pushes the prosthetic arm; the prosthetic arm collides into an unseen obstacle) that may occur in the environment [27, 28].

It must be noted that employing adaptive control in the prosthetic control loop does not systematically guarantee a better performance and/or a reduced cognitive workload. For instance, after the study, personal interviews with the users revealed that if a target was not reached in the active mode, it was sometimes awkward to switch back to the traditional (passive) mode in order to regain control of the prosthetic arm and reach the target. This illustrates how the implementation of the synergistic control between the user and the robot is critical. In this regard, a biologically plausible neurocontroller trained on-line may provide a better user-robot functional merging. This also emphasizes the need for future works that include the development of improved switching modes, more complex tasks and enhanced bio-mimetic control systems that incorporate both kinematics and dynamics characteristics of the prosthetic device. The long term goal of this work is to develop intelligent collaborative human-robotic systems to improve rehabilitation.

References

1. Pinheiro Jr., C.G., Naves, E.L., Pino, P., Losson, E., Andrade, A.O., Bourhis, G.: Alternative communication systems for people with severe motor disabilities: a surve. Biomed. Eng. Online 10, 31 (2011)
2. Chin, C.A., Barreto, A.: Enhanced hybrid electromyogram/eye gaze tracking cursor control system for hands-free computer interaction. In: IEEE EMBS Proceedings, New York, USA, pp. 2296–2299 (2006)
3. Chen, Y.L., Kuo, T.S., Chang, W.H., Lai, J.S.: A novel position sensors-controlled computer mouse for the disabled. In: IEEE EMBS Proceedings, Chicago, USA, pp. 2263–2266 (2000)
4. Choi, C., Kim, J.: A real-time EMG-based assistive computer interface for the upper limb disabled. In: IEEE ICORR Proceedings, Noordwijk, The Netherlands, pp. 459–462 (2007)
5. Evans, D.G., Drew, R., Blenkon, P.: Controlling mouse pointer position using an infrared head-operated joystick. IEEE Trans. on Rehabilitation Engineering 8(1), 107–117 (2000)
6. Huo, X., Wang, J., Ghovanloo, M.: A magneto-inductive sensor based wireless tongue-computer interface. IEEE Trans. on Neural Systems and Rehabilitation Engineering 16, 497–503 (2008)
7. Perez-Maldonado, C., Wexler, A.S., Joshi, S.S.: Two-dimensional cursor-to-target control from single muscle site sEMG signals. IEEE Trans. on Neural Systems and Rehabilitation Engineering 18, 203–209 (2010)
8. Williams, M.R., Kirsch, R.F.: Evaluation of head orientation and neck muscle EMG signals as command inputs to a human-computer interface for individuals with high tetraplegia. IEEE Trans. on Neural Systems and Rehabilitation Engineering 16, 485–496 (2008)
9. Wolpaw, J.: Brain-computer interfaces as new brain output pathways. J. Physiol. 579(3), 613–619 (2007)
10. Sanchez, J.C., Mahmoudi, B., DiGiovanna, J., Principe, J.C.: Exploiting co-adaptation for the design of symbiotic neuroprosthetic assistants. Neural Netw. 22(3), 305–315 (2009)
11. Pilarski, P.M., Dawson, M.R., Degris, T., Fahimi, F., Carey, J.P., Sutton, R.S.: Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In: IEEE ICORR Proceedings 2011, p. 5975338 (2011)

12. Abascal, J.: Users with disabilities: maximum control with minimum effort. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2008. LNCS, vol. 5098, pp. 449–456. Springer, Heidelberg (2008)
13. Zeng, Q., Teo, C.L., Rebsamen, B., Burdet, E.: A collaborative wheelchair system. IEEE Trans. on Neural Systems and Rehabilitation Engineering 16(2), 161–170 (2008)
14. Carlson, T., Demiris, Y.: Collaborative control for a robotic wheelchair: evaluation of performance, attention, and workload. IEEE Trans. Syst. Man Cybern. B Cybern. 42(3), 876–888 (2012)
15. Bullock, D., Grossberg, S., Guenther, F.H.: A self organizing neural model for motor equivalent reaching and tool use by a multijoint arm. J. Cog. Neurosc. 5(4), 408–435 (1993)
16. Guenther, F.H., Micci-Barreca, D.: Neural models for flexible control of redundant systems. In: Morasso, P.G., Sanguinetti, V. (eds.) Self-Organization, Computational Maps and Motor Control, pp. 383–421. Elsevier, Psychol. series, The Netherlands (1997)
17. Srinivasa, N., Bhattacharyya, R., Sundareswara, R., Lee, C., Grossberg, S.: A bio-inspired kinematic controller for obstacle avoidance during reaching tasks with real robots. Neural Netw. 35, 54–69 (2012)
18. Gentili, R.J., Oh, H., Molina, J., Contreras-Vidal, J.L.: Neural Network Models for Reaching and Dexterous Manipulation in Humans and Anthropomorphic Robotic Systems. In: Custuridis, V., Hussain, A., Taylor, J.G. (eds.) Perception-Action Cycle: Models, Architectures and Hardware. Springer Series in Cognitive and Neural systems, pp. 187–218. Springer, New York (2011)
19. Pedreño-Molina, J.L., Molina-Vilaplana, J., Coronado, J.L., Gorce, P.: A modular neural network linking Hyper RBF and AVITE models for reaching moving objects. Robotica 23(5), 625–633 (2005)
20. Gentili, R.J., Oh, H., Molina, J., Contreras-Vidal, J.L.: Cortical network modeling for inverse kinematic computation of an anthropomorphic finger. In: IEEE EMBS Proceedings, Boston, USA, pp. 8251–8254 (2011)
21. Gentili, R.J., Oh, H., Molina, J., Reggia, J.A., Contreras-Vidal, J.L.: Cortex inspired model for inverse kinematics computation for a humanoid robotic finger. In: IEEE EMBS Proceedings, Boston, USA, pp. 3052–3055 (2012)
22. Kitazawa, S., Goto, T., Urushihara, Y.: Quantitative evaluation of reaching movements in cats with and without cerebellar lesions using normalized integral of jerk. In: Mano, N., Hamada, I., DeLong, M. (eds.) Role of the Cerebellum and Basal Ganglia in Voluntary Movement Amsterdam, pp. 11–19. Elsevier, The Netherlands (1993)
23. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human Mental Workload 1, 139–183 (1988)
24. Huynh, H., Feldt, L.S.: Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational and Behavioral Statistics 1, 69–82 (1976)
25. Royer, A.S., He, B.: Goal selection versus process control in a brain-computer interface based on sensorimotor rhythms. J. Neural Eng. 6(1), 016005 (2009)
26. Royer, A.S., Rose, M.L., He, B.: Goal selection versus process control while learning to use a brain-computer interface. J. Neural Eng. 8(3), 036012 (2011)
27. Miller, M.W., Rietschel, J.C., McDonald, C.G., Hatfield, B.D.: A novel approach to the physiological measurement of mental workload. Int. J. Psychophysiol. 80(1), 75–78 (2011)
28. Rietschel, J.C., Miller, M.W., Gentili, R.J., Goodman, R.N., McDonald, C.G., Hatfield, B.D.: Cerebral-cortical networking and activation increase as a function of cognitive-motor task difficulty. Biol. Psychol. 90(2), 127–133 (2012)

Combining Augmented Cognition and Gamification

Curtis S. Ikehara[1], Martha E. Crosby[1], and Paula Alexandra Silva[1,2]

[1] University of Hawaii at Manoa, Department of Information and Computer Sciences
1680 East-West Rd., POST 317, Honolulu, HI 96822
[2] Universidade Portucalense Infante D. Henrique, Departmento de Inovação Ciência e Tecnologia, Rua Dr. António Bernardino de Almeida, 541, 4200-072 Porto
{cikehara,crosby,paulaale}@hawaii.edu, palexa@gmail.com

Abstract. The strategic goal of augmented cognition is to increase task performance capacity by using physiological sensor feedback to adjust or modify the activity for the user. Gamification has been shown to increase performance by using certain combinations of game elements. Both augmented cognition and gamification address increased task performance capacity. Gamification adds to augmented cognition by directly addressing the motivation of the user to remain engaged in the activity. This has also been referred to as flow, or the optimal experience. This paper describes an example of a gamified activity in which the physiological sensors of augmented cognition are used to foster the optimal experience desired in gamification. Also, discussed is how the strategic goals of augmented cognition and gamification overlap through the use of a gamified example that describes how the components of augmented cognition and elements of gamification can be used together to better achieve the goal of increased task performance capacity.

Keywords: augmented cognition, gamification, physiological sensors.

1 Introduction

The phenomenon of gamification has been gathering a great deal of interest among the various quadrants of society. It has been applied to a variety of areas, such as business[1], education[2], and health[3], and it has been included in the Gartner [1] hype cycle for emerging technologies in the last two years

There have been several attempts to define gamification. Deterding et al. [2] have defined it as "the use of game design elements in non-game contexts". Werbach and Hunter [3] (p. 26) have elaborated on the definition to be "the use of game elements and game-design techniques in non-game contexts". From these two definitions it becomes clear that gamification is not about building a fully fledge game, but rather about using parts of one. Deterding et al. [2] clarifies that these are "elements that are found in most (but not necessarily all) games, readily associated with games, and

[1] Foursquare - https://foursquare.com/
[2] Chore Wars -http://www.chorewars.com/
[3] Fitocracy - https://www.fitocracy.com/

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 676–684, 2013.

found to play a significant role in game play". Gamification leverages elements of games to promote users' motivation and to create engaging dynamics that can eventually influence and/or change the user's behavior.

Games are generally regarded as enjoyable and fun, but most interestingly, they have shown to motivate users to engage with them with unparalleled intensity and duration[2]. Moreover, research into human motivation demonstrates that people feel motivated by well-designed game features [3] (p. 10). It is this compelling nature of games that gamification researchers want to explore and capitalize in order to improve the effectiveness in other areas. This paper is particularly interested in how both augmented cognition and gamification can increase task performance capacity in education. Section 2 describes an example of a prototypical gamified activity and discusses how combining augmented cognition and gamification can support task performance capacity.

1.1 Background

Flow has been associated both with games and education. Flow, or the optimal experience, as described by Csikszentmihalyi refers to "a sense of that one's skills are adequate to cope with the challenges at hand in a goal directed, rule bound action system that provides clear clues as to how one is performing. Concentration is so intense that there is no attention left over to think about anything irrelevant or to worry about problems. Self-consciousness disappears, and the sense of time becomes distorted." [4] (p. 71).

The quality of an experience depends on an individual's level of challenge and skill when performing a given activity. Optimal experiences, i.e.: flow experiences, are likely to occur when both skills and challenges are high, when a person's skills are fully involved in overcoming a challenge that is just about manageable [5]. The repetition of flow moments will form a narrow flow channel (Fig. 1) within which the individual is in the desirable and enjoyable state of flow. Ideally, in order to excel and deeply engage in a given activity the individual's state should be located within this channel.

Falstein [6] studied the concept of Flow in fun and games to explain that game difficulty should vary in waves. Ikehara and Crosby [7] also recognized that in augmented cognition maintaining flow within an optimum cognitive load range would enhance learning. Ideally, if we were able to systematically adjust the level of challenge and skill a user faces we could hypothetically foster the efficacy of learning.

According to Csikszentmihalyi [4] (p. 71, 72), ". . . it is much more likely that flow will result either from a structured activity, or from an individual's ability to make flow occur, or both", but activities can also be designed to make optimal experiences easier to occur. It is easy to enter flow in games and these are actually exemplar flow activities. Important to that is the existence of clear goals and rules and of immediate feedback. Goals make it possible to act without thinking while rules direct energy in patterns that are enjoyable [4] (p. 76). Finally, immediate feedback makes the person aware of how well she is doing and enables the person to know whether she improved her position or not after each move.

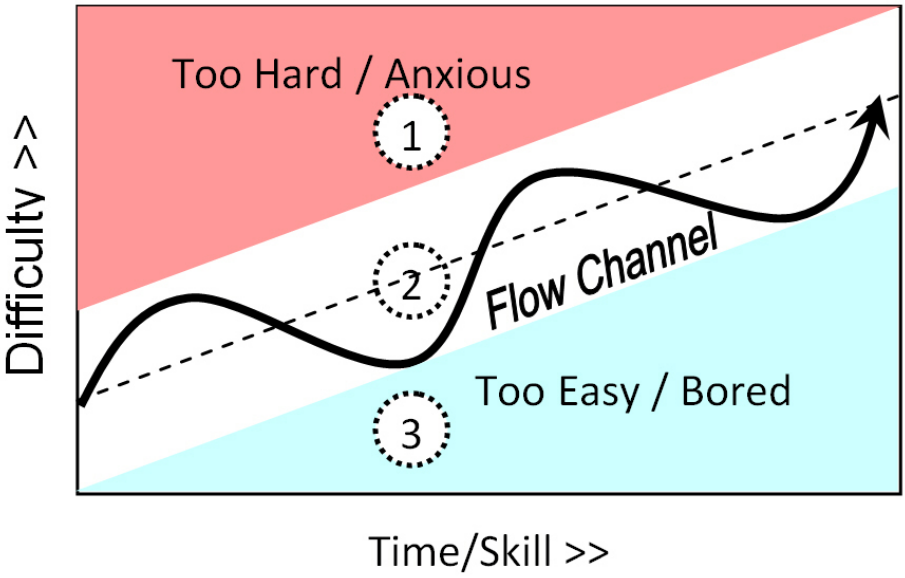


Fig. 1. The flow channel based on Csikszentmihalyi [4] (p. 74) and Falstein [6] depicting numbered example states

Due to the common complex combination of elements that are present in games it is difficult to exactly pinpoint the game elements and relationships between them that contribute to the optimal enjoyable experiences. Nonetheless, it is known that certain combinations of game elements enable flow to occur.

Researchers have tried to unfold those properties of games that make them so compelling. Jane McGonigal identified four game traits: goals, rules, feedback, and voluntary participation [8] (p. 21). Hunicke et al. [9] pulled out the components of games by formalizing them in the MDA framework that includes, mechanics, dynamics and aesthetics. Reeves and Read [10] identified the ten ingredients of great games: self representation with avatars, three-dimensional environments, narrative context, feedback, reputations, ranks, and levels, marketplaces and economies, competition under rules that are explicit and enforced, teams, parallel communication systems that can be easily configured, and time pressure. Werbach and Hunter [3] (p. 77-82) propose a pyramid of elements that from the top-down includes three levels of game elements: dynamics, mechanics and components (Table 1). These are used later in the description and discussion of how augmented cognition can be applied to a prototypical gamified activity to increase task performance.

Increasing task performance capacity is at the core of both gamification and augmented cognition. Augmented cognition "aims at evaluating in real-time the cognitive state of a user (e.g. EEG), and to design closed-loop systems to modulate information flow with respect to the user's cognitive capacity." [11] In order to increase the learning rate, the ability to do a task, or to maintain continued competent task performance,

Table 1. Werbach and Hunter Pyramid of Elements [3]

High Level - Dynamics	
H01	Constraints (that trigger meaningful choices)
H02	Emotions (what can make the experience richer and whatever motivates the people to play more)
H03	Narrative (what makes the gamified system coherent)
H04	Progression (what gives the player the sense that they are progressing towards the objective)
H05	Relationships(people interacting with each other (e.g., teams)
Mid-Level - Mechanics	
M01	Challenges (objective to reach)
M02	Chance (the luck involved)
M03	Competition (getting people to compete against each other)
M04	Cooperation (getting people to work together)
M05	Feedback (what enables the users to see how they are doing in real time and tends to drive them along to go further)
M06	Resource Acquisition (the things that the game gives you opportunity to get in the game in order to move it forward)
M07	Rewards (some benefits that you get for some achievement in the game)
M08	Transactions (buying and selling, or exchanging something with other players, or with what's called a non player character, with some automated character in the game)
M09	Turns (the opportunity or obligation to do something that comes successively to each of a number of people)
M10	Win States - The state which defines winning the game
Bottom Level - Components	
B01	Achievements (defined objectives)
B02	Avatars (visual representations of a player's character)
B03	Badges (visual representation of achievements)
B04	Boss fights (especially hard challenges and the culmination of a level)
B05	Collections (set of items or badges to accumulate)
B06	Combat (a defined battle, typically short-lived)
B07	Content unlocking (aspects available only when players reach objectives)
B08	Gifting (opportunities to share resources with others)
B09	Leaderboards (visual displays of player progression and achievements)
B10	Levels (predefined steps in player progression)
B11	Points (numerical representations of game progression)
B12	Quests (predefined challenges with objectives and rewards)
B13	Social Graphs (representation of players' social network within the game)
B14	Teams (defined groups of players working together for a common goal)
B15	Virtual goods (game assets with perceived or real-money value)

information is obtained from a variety of physiological sensors and used to adjust or modify a task. Elements of games have been used in augmented cognition to enhance task performance on an *ad hoc* basis [12]. The array of game elements that is used in gamification can be systematically leveraged to enhance augmented cognition systems.

2 Description of a Prototypical Gamified Activity

For children, learning fractions can be an increasingly challenging task as the numerator and denominator of the fraction increase in size. The strategic goal of teaching fractions is accomplished by getting children to repetitively practice fraction problems as the difficulty increases. A typical fraction exercise would consist of the set of fraction problems shown below.

Is $\frac{1}{2} > \frac{1}{3}$?, Is $\frac{2}{9} > \frac{1}{3}$?, Is $\frac{4}{9} > \frac{1}{3}$?, Is $\frac{11}{18} > \frac{1}{3}$?

To meet the strategic goal of the child learning fractions and the child's goal of participating in an engaging activity the fraction task can have game elements incorporated.

The Moving Targets Fractions (MTF) activity, described by Ikehara, Chin and Crosby [13] is a gamified version of the typical fraction activity shown above. The MTF task presents a fixed number of oval targets containing fractions on a computer screen. These fractions float across the screen from left to right (see Figure 2). The primary goal of the user is to maximize the score by selecting the fractions greater than 1/3 before they reach the right edge of the screen.

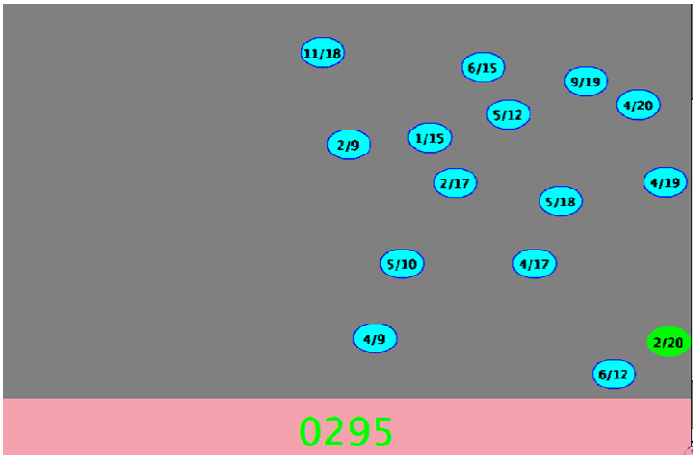


Fig. 1. Screen capture of the Moving Targets Fraction (MTF) task

There are several goals and subgoals in this activity, but interwoven into these are the elements of gamification used to motivate the user to achieve the strategic goals. Note that the game element code from Werback Gamification Pyramid of Elements

follows element description. For example, high score (M07, B11) refers to high score being related to 'Points' from Werbach and Hunter Gamification Pyramid.

For the user to obtain the highest score (H02, H03, M01, M07, B11), the user must select all fractions greater than the critical value of 1/3 before they touch the right edge of the screen (H01). The goal of the user is to maximize the score (M01, M07), which is prominently displayed at the bottom of the screen (M05, B09), by achieving four subgoals before taking action.

- The first subgoal is to evaluate all fractions as they appear to determine if the fraction can be evaluated quickly or with difficulty (M01).
- The second subgoal is to evaluate each fraction's value, within the user's confidence level, to determine its relationship to a critical value (M01). Fractions greater than the critical value will increase the score when selected (B11). The difficulty of the comparison (i.e., cognitive load) is controlled by the fraction selected. For example, 1/2 is obviously greater than 1/3, but comparing 6/17 versus 1/3 requires much more cognitive effort. The user registers decisions by clicking with the mouse on those fraction targets greater than the critical value. Correctly selected fraction bubbles turn green. Incorrectly selected fraction bubbles turn red (M05). For even greater difficulty, the critical value can be changed from simple fractions like 1/3 to complex fractions such as 5/13 (M01).
- The third subgoal is to consider how the score is computed when selecting targets. The score is computed as 100 times the fractions that the user selects correctly above 1/3 (e.g., 3/4 * 100 = +75) and deducts 100 points for each incorrect selection. The negative scores for incorrect targets means the user cannot simply select everything on the screen to maximize the score (H01, M01).
- The fourth subgoal is to not let a fraction greater than the critical value touch the right edge. This fraction bubble will turn red. A deduction of 200 times the fraction value will occur if a fraction greater than the critical value touches the right side of the screen while a score of 200 point are added when the fraction is below the critical value. This motivates the subject to evaluate all fractions presented and not just the easily computed ones (H01, M01).

The subgoals can take on different priorities depending on task variables such as the difficulty of evaluating the fraction, the value of the fraction, how close the fraction is to the right side of the screen and the number of fractions presented (H01, M01). The priorities of the subgoals can also be affected by user factors such as arousal, stress and motivation (H02).

With augmented cognition, incorporating physiological sensors to adjust or modify game elements it is possible to keep the user in the optimum flow channel is possible by identifying two general classes of easily modifiable game elements: challenges and rewards. Modifiable challenges of the game includes: win states, time pressure, chance, transactions, content unlocking and quests. Rewards includes: leaderboards, reputations, ranks, levels, resource acquisition, badges, collections, points and virtual goods. Both challenges and rewards can be modified based on physiological sensor to keep the user in the flow channel. See Table 2 for a list of physiological sensors, physiological measures and secondary measures than can be used to direct the modification of game elements.

Table 2. Sensors, physiological measures and secondary measures

Sensors	Physiological Measures	Secondary Measures
Eye Position Tracker	Gaze Position, Fixation Number, Fixation Duration, Repeat Fixations, Search Patterns	Difficulty, Attention, Stress, Relaxation Problem Solving, Successful Learner, Higher Level of Reading Skill [14], [15]
	Pupil Size, Blink Rate, Blink Duratio	Fatigue, Difficulty, Strong Emotion, Interest, Mental Activity - Effort, Familiar Recall, Positive / Negative Attitudes, Information Processing Speed [14]
Mouse Pressure	Pressures Applied to the Mouse Case and Buttons.	Stress, Certainty of Response, Cognitive Load [16], [17]
Skin Conductivity	Tonic and Phasic Changes	Arousal [14]
Temperature	Finger, Wrist and Ambient Temperature	Negative Affect (Decrease), Relaxation (Increase) [14]
Relative Blood Flow	Heart Rate and Beat to Beat Heart Flow Change	Stress, Emotion Intensity [14]

Three examples below describe the 'task state', 'physiological sensors used', 'user state detected by the sensors' and 'element of the game modified' (see Table 3). Refer to Figure 1 for where these three examples are located in relation to the flow channel.

Table 3. Examples of physiological sensor directed modification of game elements

	Task State	Physiological Sensors Used	User State Detected by the Sensors	Element of the Game Modified
	Fractions are too difficult	Relative blood flow - heart rate above normal	Sensors indicate a persistent high level of arousal, the person is above the flow channel boundary	Reduce challenge by reducing the time pressure by slowing the flow of fractions or by reducing the level of difficulty of the fraction
	Fraction difficulty is appropriate	Skin conductivity-normal	Indicate a normal level of arousal, the person is within the flow boundary	No change
	Fractions are too simple	Eye Tracking - Blink duration longer than normal	Indicate a persistent low level of arousal, the person is below the flow channel boundary	Challenge modification is to increase the time pressure or increase the level of difficulty to move the user to an increased skill level

As shown in the example above, information from the physiological sensors can be used to modify game elements in the activity. This allows the achievement of the strategic goal of teaching fractions, while maintaining the positive attitude of the individual to continue performing the activity. Concurrent modification of challenges and rewards can be used to maintain or move a user higher in the flow channel. For example, increasing the time pressure (i.e., Challenge) while increasing the rank (i.e., Reward) could be used to move a user to a higher level of skill or challenge.

3 Conclusion

There is an alignment of the strategic goals of augmented cognition and gamification. This paper describes an example of an activity in which gamification elements are found in augmented cognition activities to increase task performance capacity. The Moving Targets Fractions task provides an example where game elements are identified and the potential of modifying game elements based on physiological sensors is described. Established are the relationships between the physiological sensors of augmented cognition and elements of gamification and how using sensor information to direct the modification of game elements can lead to achieving the strategic goal of increase task performance capacity while motivating the individual to achieve a high level of competence.

References

1. Gartner Research, Gartner's 2012 Hype Cycle for Emerging Technologies Identifies "Tipping Point" Technologies That Will Unlock Long-Awaited Technology Scenarios (August 16, 2012), http://www.gartner.com/newsroom/id/2124315 (accessed February 20, 2013)
2. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining gamification. In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (2011)
3. Werbach, K., Hunter, D.: For the Win: How Game Thinking Can Revolutionize Your Business. Wharton Digital Press, Philadelphia (2012)
4. Csikszentmihalyi, M.: Flow: the psychology of optimal experience. Harper & Row, Publishers, Inc. (1990)
5. Csikszentmihalyi, M.: Finding Flow: The Psychology of Engagement with Everyday Life. HarperCollins Publishers, Inc., Perennial (1997)
6. Falstein, N.: Understanding Fun—The Theory of Natural Funativity. In: Introduction to Game Development, Boston, pp. 71–98 (2005)
7. Ikehara, C.S., Crosby, M.E.: Real-Time Cognitive Load in Educational Multimedia. In: World Conference on Educational Multimedia, Chesapeake (2003)
8. McGonigal, J.: Reality is broken: Why games make us better and how they can change the world. Penguin Press HC (2011)
9. Hunicke, R., LeBlanc, M., Zubek, R.: MDA: A formal approach to game design and game research. In: AAAI Workshop on Challenges in Game AI (2004)
10. Reeves, B., Read, J.L.: Total Engagement: How Games and Virtual Worlds Are Changing the Way People Work and Businesses Compete. Harvard Business Press (2009)

11. Augmented cognition - Wikipedia, http://en.wikipedia.org/wiki/Augmented_cognition (accessed February 20, 2013)
12. Crosby, M.E., Ikehara, C.S.: Using Physiological Measures to Identify Individual Differences in Response to Task Attributes. In: Foundations of Augmented Cognition, 2nd edn., pp. 162–167. Strategic Analysis, Inc., San Ramon (2006)
13. Ikehara, C.S., Chin, D.N., Crosby, M.E.: A model for integrating an adaptive information filter utilizing biosensor data to assess cognitive load. In: User Modeling (2003)
14. Andreassi, J.: Psychophysiology: Human Behavior and Physiological Response, 3rd edn. Lawrence Erlbaum, New Jersey (1995)
15. Sheldon, E.: Virtual agent interactions. University of Central Florida, Orlando (2001)
16. Ikehara, C.S., Crosby, M.E., Chin, D.N.: A Suite of Physiological Sensors for Assessing Cognitive States. In: 11th International Conference on Human-Computer Interaction (2005)
17. Ikehara, C.S., Crosby, M.E.: Assessing Cognitive Load with Physiological Sensors. In: 38th Annual Hawaii International Conference on System Sciences (2005)

Issues in Implementing Augmented Cognition and Gamification on a Mobile Platform

Curtis S. Ikehara, Jiecai He, and Martha E. Crosby

University of Hawaii at Manoa, Department of Information and Computer Sciences,
1680 East-West Road, Honolulu, Hawaii 96822, USA
{cikehara,jiecai,crosby}@hawaii.edu

Abstract. There are two major trends in computing that will impact augmented cognition. The first is the shift in computing platform from the desktop to mobile computing (e.g., smartphone and tablet) because the user wants to be able to do computing tasks where ever they are. The second trend is the gamification of computer applications to keep the user engaged and motivated. Compared to a workstation, the mobile computing environment is a challenge because of limited computing power, storage capacity, internet connectivity and battery capacity. This paper discusses the issues involved in implementing augmented cognition activities on a mobile platform and the tradeoffs of gamifying augmented cognition activities. These issues are discussed in terms of two example mobile platform applications that implement internal and external sensors.

Keywords: mobile computing, augmented cognition, gamification, physiological sensors.

1 Introduction

There are two major trends in computing that will impact augmented cognition. The first is the shift in computing platform from the desktop to mobile computing (e.g., smartphone and tablet). "Mobile internet usage is predicted to overtake desktop usage as early as 2014."[1] Users enjoy the portability of mobile computing (e.g. smartphones and tablets). With 4G and Wi-Fi hot spots internet connectivity is almost ubiquitous. The second trend is gamification. Gamification is the incorporation of game elements into non-game applications to keep the user engaged and motivated.

1.1 Mobile Computing

With the availability of high speed wireless internet (e.g., 4G or Wi-Fi) and cloud computing the computing capacity of the mobile user has significantly increased. Organizations with traditional web services are becoming increasingly aware of the need to migrate or redesign their applications to the mobile computing platform. One of the primary concerns with the shift to mobile computing will be how to maintain and increase user productivity.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 685–694, 2013.

1.2 Augmented Cognition

Maintaining and increasing user productivity, which is to increase task performance capacity, is a strategic goal of augmented cognition. This task performance capacity could be manifested by increasing the learning rate, increasing the ability to do a task, or maintaining continued task competence. In augmented cognition, increase task performance capacity is achieved by using physiological sensor feedback to adjust or modify the activity the user is performing.

To implement a real-time augmented cognition system on a workstation can be a challenge because of the streaming physiological sensor data that must be stored and processed while simultaneously running and modifying the application. Current mobile platforms in comparison are more limited in computing and storage capacity than the workstation and must also consider limited battery life. Regardless of these limitations of the mobile systems, computing power and storage increases seem to be following Moore's law [2]. Also, with the advent of higher wireless communication rates, both computing power and storage capacity could be off-loaded to the cloud. Preliminary research and methods developed now can be applied to future mobile devices with greater computing power, storage capacity, wireless speed and battery life.

Table 1. Mobile computing device sensors, physiological measures and potential cognitive measures

Sensors	Physiological Measure	Potential Cognitive Measure
Camera	Eyetracking : Gaze Position, Fixation Number, Fixation Duration, Repeat Fixations, Search Patterns, Pupil Size, Blink Rate, Blink Duration	Difficulty, Attention, Stress, Relaxation, Problem Solving, Successful Learner, Higher Level of Reading Skill [3] [4]
	Facial Recognition	Happiness, Sadness, Surprise, Anger, Disgust, and Fear [5]
Accelerometer, gyroscope, compass	Body Motion	Arousal [3]
Touch Screen	Pressures Applied to the Button	Stress, Certainty of Response, Cognitive Load [6], [7]
Microphone	Voice Characteristics	Depression [8]

At the core of augmented cognition research to increase task performance capacity is physiological sensor selection, data collection, data analysis, and then the modification of the activity guided by the analysis of the sensor data. Currently, mobile devices are equipped with a set of sensors that could be repurposed for augmented cognition (see Table 1). An eye-tracking application using the forward facing camera on an Android based tablet is described in this paper. Other sensors found on mobile devices are listed in Table 1 along with their potential cognitive measures that could be used with augmented cognition applications.

Almost all mobile devices have the option to have Bluetooth, Wi-Fi and 4G communication. Assuming Wi-Fi or 4G may be in use accessing cloud computing processing and data storage resources, Bluetooth would be the preferred interface approach for connecting sensors in close proximity. In this paper, an Android smart phone application is described to demonstrate how Bluetooth would be implemented on a mobile device to acquire sensor data.

1.3 Gamification

Gamification is a process of applying game elements to non-game applications to maintain a high level of user engagement and motivation to influence behavior.

Jane McGonigal identified four game traits: goals, rules, feedback, and voluntary participation [9] (p. 21). Werbach and Hunter [10] (p. 77-82) proposed a large list of game elements broken down into three levels of elements: dynamics, mechanics and components. Reeves and Read [11] identified the ten ingredients or game elements of great games (Table 2). Game elements place demands on computing, data storage, and wireless connectivity. Table 2 used Reeves and Read's list and is not exhaustive, but provides linkages between game elements and the resources that could be needed. Note that all resources (i.e., computing, storage, connectivity and sensors) reduce battery life. In the mobile environment, gamification has several benefits that must be weighed against the drawbacks.

Gamification Benefits

- A high level of user engagement and motivation.
- Gamification can be enhanced by tapping into the physiological sensors used in augmented cognition.
  - Sensor data can be used to direct real-time feedback to the user.
  - Adjustments to rewards and the difficulty of the activity can be based on the cognitive state derived from sensor data. Currently, gamification relies on the user actions or overall performance to adjust rewards and activity difficulty.

Table 2. "Ten Ingredients of Great Games" Reeves and Read [11] and resources used

	Ingredients (Game Elements)	Computing	Data Storage	Wireless Connectivity	Sensors
1	Self representation with avatars	To display of the avatar.	To store multiple avatars.		
2	Three-dimensional environments	To display the environment.	To store the environment.		
3	Narrative context		To store the narrative.		To adjust the challenge.
4	Feedback	To ascertain performance and provide feedback.	To record performance history.		To adjust feedback.
5	Reputations, ranks, and levels		To record reputations, ranks, and levels.	To display for other players	To adjust the rewards.
6	Marketplaces and economies		To store marketplace information.	To display for other players.	
7	Competition under rules that are explicit and enforced			To transfer information between competitors.	
8	Teams			To transfer information between collaborators.	
9	Parallel communication systems that can be easily configured			To provide communications between players.	
10	Time pressure.	To run the task at a rapid pace.			

Gamification Drawbacks

- Adding game elements could degrade the performance of the non-game activity being gamified. Sensor data could overwhelm limited computing, storage and connectivity resources. Battery life could also be significantly reduced. In some cases, to increase battery life, wireless connectivity to the cloud could be used to supplement or replace computing and data storage.
- Extensive testing may be required to fine tune the gamification of an activity. It is unclear which combination of game elements produces the best results and

individual differences may play a significant role in determining the optimum combination of game elements. Also, the combination of game elements may need to dynamically change with the individual's predisposition.

2 Examples of Sensors for Mobile Applications

2.1 Using the Front Facing Camera of a Mobile Device for Eye Tracking

Many smartphones and tablets have forward facing cameras (i.e., cameras that face the user) for video conferencing. Although these cameras have a lower resolution than the rear facing camera they are of sufficient quality to do eye-tracking. Although the pupil-center/corneal-reflection eye tracking technique using both the pupil location and a reflected glint from the eye is more accurate, the less accurate pupil-center only approach is possible with a mobile device.

The first step of locating the two pupils of the user's eyes begins with capturing and image of the user facing the display. The second step is to locate the face. The third step is to locate the eyes. The fourth step is to locate the pupils of both eyes. A calibration procedure is required for each user where the user looks at specific locations on the screen and the pupil location is recorded. Once that calibration information is available, an algorithm can be used to determine the rough location of the gaze of the user. The following gives more detail on the eye-tracking process using the camera image.

An Android tablet (Eee Pad Transformer TF101) was programmed based on the OpenCV class for face detection using JAVA as the programming language. "OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library" [12]. A Local Binary Patterns (LBP) cascade classifier is used to do face detection. "LBP features are integer in contrast to Haar features, so both training and detection with LBP are several times faster than with Harr features." [13] Algorithm efficiency is critical with a mobile device since an efficient algorithm would increase the speed of computational and reduced power consumption.

To locate the eyes in the face a Harr cascade classifier is used. Image processing ". . . detectors based on these Haar-like features work well with 'blocky' features such as eyes, mouth, face, and hairline . . . " [14] (p. 510).

Locating the pupils is more complicated. There are many issues including eyelids, eyelashes, corneal reflections, shadows, and blinking. Having an algorithm to deal with all these factors is beyond the scope of this paper, but researchers have been working on these problems. Even with a less robust pupil detection method used, the pupil is difficult to locate and requires several image processing steps. Figure 1 shows the end result of the image processing steps below.

- Turn the color image containing the eye into gray scale.
- Eliminate unnecessary pixels above the eye area.
- Histogram Equalization to increase black and white contrast to make the black pupil the dominant in terms of pixel intensity.

- Invert pixel intensity to make the black pupil white.
- Use erosion to "eat away" the distracting white areas.
- Threshold the picture into binary to make the image only black and white.
- Take the center of the bounding box of the contour as the center of the pupil.

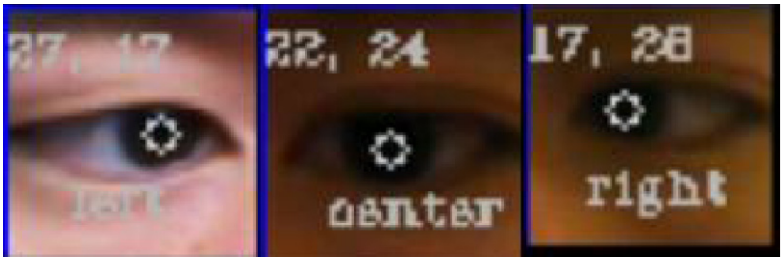


Fig. 1. The location of the pupil looking left, center and right

The horizontal location of the pupil can be determined more accurately than the vertical location because there is a clear image change when moving the eyes from left to right on the display than up and down on the display. Figure 2 shows the face location, eye location and pupil location. Note that the pupil position on the left eye is not centered since the pupil detection algorithm used becomes less accurate when there is a corneal reflection on the pupil.

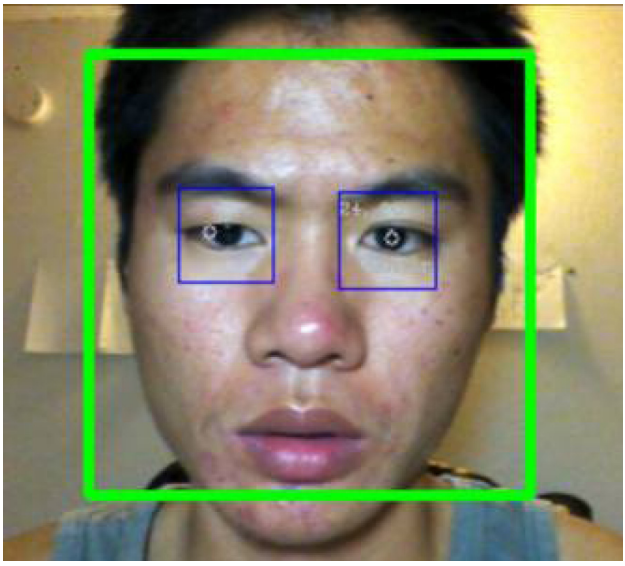


Fig. 2. Eyes and face are detected. Note that the pupil position on the left eye is not centered since the algorithm used becomes inaccurate when there is a corneal reflection on pupil.

With the pupil location accurately determined calibration of the user can be performed. Both calibration data and the data from the real-time location of the two pupils can be used to determine where the user is looking at on the screen.

2.2 A Smartphone Application Using External Sensors Connected via Bluetooth

At times, it is desirable to have sensors that are not located on the device or different sensors are needed. Described is a simple application demonstrating the potential of an external sensor connected via Bluetooth. The system consists of an Android smartphone and Bluetooth system with several sensors (see Figure 3). Bluetooth communication allows the Bluetooth system to be placed up to several feet away from the smartphone. A mathematics game for children written in JAVA based on comparing fractions is implemented on the smartphone and light sensors connected through Bluetooth are used to indicate the relationship between the two fractions (i.e., larger, smaller or equal).

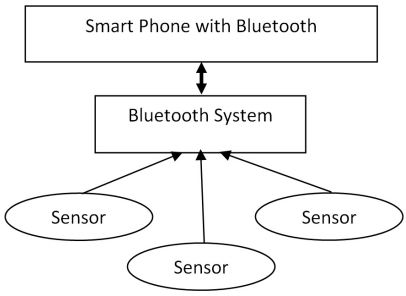


Fig. 3. The system consists of an Android smartphone, Bluetooth link and several sensors

An Arduino UNO with Bluetooth (i.e., the Bluetooth system) is connected to three Arduino Pro Minis connected to light sensors (see Figure 4). Inter-Integrated Circuit (I2C), a two wire protocol, is used to allow the Arduino devices to communicate with each other.

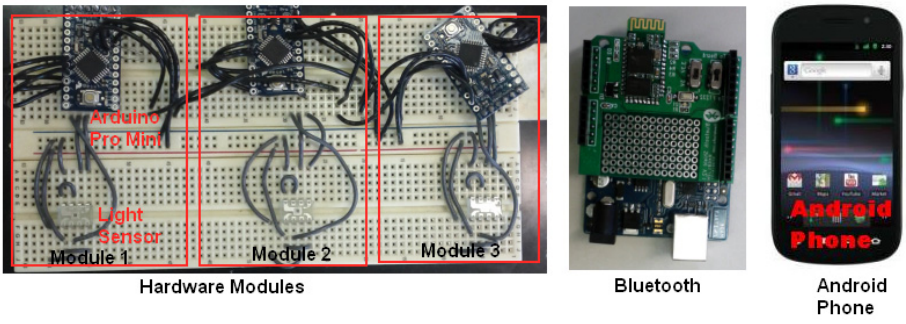


Fig. 4. Complete system with sensor hardware, processor with Bluetooth and Android phone

The power is turned on for the Bluetooth system with sensors then the smartphone is set to search for Bluetooth devices. Once the devices are linked the fraction game can begin. The fraction game shows two fractions and asked the player to determine if the first fraction is larger, smaller or equal to the bottom fraction (see Figure 5). The player blocks the appropriate light sensor in response to the displayed question.

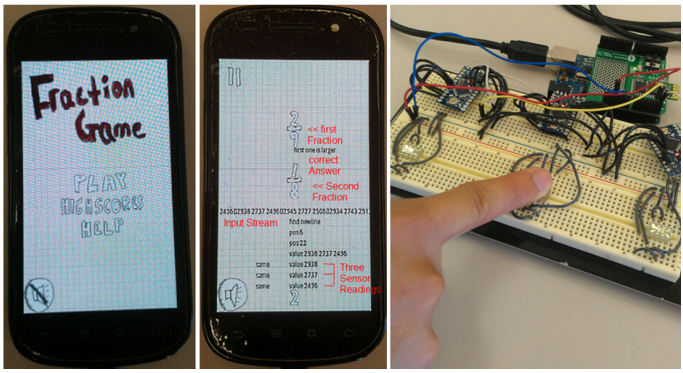


Fig. 5. Left - Prototype display. Center - Prototype display with fractions and debugging information. Right - The player is blocking the light sensor to indicate that the fraction values are equal.

This system uses the I2C communications protocol allowing any number of sensors and indicators (i.e. LED, lights, motors) can be connected to this system. The I2C communications protocol is also used by the Plug-n-Play Wearable Computing Framework [15] which has sensors and indicators integrated into clothing.

The Arduino UNO with Bluetooth also has processing and storage capacity. This processing and storage capacity can be used to reduce mobile device resource requirements.

Bluetooth Issues

There are several versions of Bluetooth. The current version 4.0 is becoming more common on current mobile computing devices. Bluetooth version 2.0 has a ". . . theoretical maximum useful data transfer rate of approximately 2.1 Megabits per second (Mbps)." [16] Bluetooth version 3.0 + HS has a data transfer speed of up to 24 Mbit/s over a collocated 802.11 link. Bluetooth version 4.0 includes Classic Bluetooth (version 1 & 2), Bluetooth high speed (version 3) and Bluetooth low energy protocols. A Bluetooth version 4 device that implements only the low energy protocol may not be compatible with earlier Bluetooth versions.

Besides version, Bluetooth uses a variety of different protocol stacks to exchange data. What this means is that both mobile and sensor device must support the same protocol stack. For example, the Apple iPhone 4 with Bluetooth v4.0 supports these protocols: A2DP, AVRCP, HFP, HID, MAP, PAN, and PBAP.

3 Discussion

Using internal or external sensors on mobile devices, although not trivial as demonstrated with eye-tracking, can be done. These sensors systems can be used to support augmented cognition on mobile devices. Doing augmented cognition on a mobile device has many challenges. These challenges relate primarily to limited computing power, storage capacity, internet connectivity and battery capacity. Gamification has the benefit of motivating the user and improving performance by appropriately inserting game elements into the mobile application, but gamification consumes resources that can exacerbate the mobile device challenges. The mobile computing platform is a fundamentally different computing experience than the workstation experience since mobile computing can occur in almost any environment. Understanding this fundamental difference is why research on mobile computing devices implementing augmented cognition and gamification, though challenging, needs to move forward.

References

1. P. UK, Mobile internet usage to overtake desktop as early as 2014 says new marketing report, Report Buyer (November 23, 2012), http://news.yahoo.com/mobile-internet-usage-overtake-desktop-early-2014-says-080101013.html (accessed February 27, 2013)
2. Chang, Y.S., Lee, J., Jung, Y.S.: Are Technology Improvement Rates of Knowledge Industries Following Moore'S Law?-An Empirical Study of Microprocessor, Mobile Cellular, And Genome Sequencing Technologies. KDI School of Pub Policy & Management Paper (2012)
3. Andreassi, J.L.: Psychophysiology: Human Behavior and Physiological Response. Lawrence Erlbaum, Hillsdale (1995)
4. Sheldon, E.: Virtual Agent Interactions. University of Central Florida, Orlando (2001)
5. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, F., Movellan, J.: The computer expression recognition toolbox (CERT). In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (2011)
6. Ikehara, C.S., Crosby, M.E.: Assessing Cognitive Load with Physiological Sensors. In: 38th Annual Hawaii International Conference on System Sciences (2005)
7. Ikehara, C.S., Crosby, M.E., Chin, D.N.: A Suite of Physiological Sensors for Assessing Cognitive States. In: Proceedings of the 1st International Conference on Augmented Cognition, Las Vegas, NV (2005)
8. Chang, D.F.K.H., Canny, J.: Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. In: Proceedings of PhoneSense 2011 (2011)
9. McGonigal, J.: Reality is broken: Why games make us better and how they can change the world. Penguin Press HC (2011)
10. Werbach, K., Hunter, D.: For the Win: How Game Thinking Can Revolutionize Your Business. Wharton Digital Press, Philadelphia (2012)
11. Reeves, B., Read, J.L.: Total Engagement: How Games and Virtual Worlds Are Changing the Way People Work and Businesses Compete. Harvard Business Press (2009)
12. O. D. Team, About OpenCV, http://opencv.org/about.html (accessed February 28, 2013)

13. Cascade Classifier Training, https://github.com/alexmac/alcexamples/blob/master/OpenCV-2.4.2/doc/user_guide/ug_traincascade.rst (accessed February 28, 2013)
14. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reily Media, Inc., Sebastopol (2008)
15. Ngai, G., Chan, S., Ng, V., Cheung, J.C., Choy, S.S.S., Lau, W.W.Y., Tse, J.: i*CATch: A Scalable, Plug-n-Play Wearable Computing Framework for Novices and Children. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010). ACM, New York (2010)
16. Kewney, G.: High speed Bluetooth comes a step closer: enhanced data rate approved (November 16, 2004), http://www.newswireless.net/index.cfm/article/629 (accessed February 28, 2013)

Visual Analysis and Filtering to Augment Cognition

Mathias Kölsch, Juan Wachs, and Amela Sadagic

jpwachs@purdue.edu, asadagic@nps.edu
http://movesinstitute.org/~kolsch

Abstract. We built and demonstrated a system that augments instructors' sensing abilities and augments their cognition through analysis and filtering of visual information. Called BASE-IT, our system helps US Marine instructors provide excellent training despite the challenging environment, hundreds of trainees and high trainee-to-instructor ratios, non-stop action, and diverse training objectives. To accomplish these objectives, BASE-IT widens the sensory input in multiple dimensions and filters relevant information: BASE-IT a) establishes omnipresence in a large training area, b) supplies continuous evaluation during multi-day training, c) pays specific attention to every individual, d) is specially equipped to identify dangerous situations, and e) maintains virtual vantage points for improved situational awareness. BASE-IT also augments and personalizes the after-action review information available to trainees.

This paper focuses on the automated data analysis component, how it supplements the information available to instructors, and how it facilitates understanding of individual and team performances on the training range.

Keywords: Augmented cognition, information analysis, training range instrumentation.

1 Introduction

The keystone in US Marine training is conducted at a purpose-built training range at the Marine Corps Air Ground Combat Center in Twentynine Palms, CA, which provides a realistic environment that immerses hundreds of trainees in a small town with houses, markets, road traffic and human role players. Instructors observe the non-stop 72 hour urban operation and provide performance feedback at regular intervals. This is a challenging situation for a number of reasons.

First, this training is the final and most realistic training provided to US Marines immediately before their deployment into theatre. There is immense pressure on both trainees and trainers to achieve the training objectives as it will have a tremendous impact on the performance in theatre. Techniques not taught or not taught well, as well as mistakes not caught in training might have severe and far-reaching consequences later.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 695–702, 2013.

Second, the training facility is very expensive due to many factors, necessitating efficient and effective training. Due to these pressures on expedience and performance, as well as due to accepted best training practices, training is rarely stopped to provide feedback. In fact, the more advanced scenarios train and evaluate multiple skills simultaneously, for many individuals, with little to no room to pause and discuss or correct mistakes until the after-action review.

Third, the training range is almost one square mile large, with many buildings, roads, foot and vehicle traffic, geographic features and other realistic aspects of a small town. It is impossible for instructors to always keep an eye on and to give feedback to every trainee on individual or group behavior. Instead, observations are made at crucial times and locations, and feedback is provided in short debriefing sessions.

Fourth, some aspects of individual and group behavior are very difficult to observe from a single vantage point. For example, the precise position of an individual and his head location behind cover cannot be determined accurately from just one point of view. Also, the formation of a squad that is on foot patrol in between buildings is hard to observe from just one location due to occlusions. Viewpoint limitations might cause actions to pass unobserved and evading evaluation and feedback.

Fifth, while video streams from cameras on the range are available and are being recorded, the sheer amount of data relegates their use to isolated review questions. Rather than augmenting the instructor's *cognition*, this additional information requires *additional attention*. Also, while pole-mounted pan-tilt-zoom (PTZ) cameras are available, aerial cameras are not. Hence, occlusions from buildings in a single view are common.

Sixth, due to the aforementioned constraints, some behavioral mistakes cannot be focused on during this training, as it would distract from the main training objectives. One such mistake is unintentionally pointing a weapon system towards a fellow Marine, also called flagging. Additionally, flagging is difficult to determine unless an instructor happens to be very near the occurrence and paying attention to the swift movements of trainees.

BASE-IT, short for Behavioral Analysis and Synthesis for Intelligent Training [1], was developed at the MOVES Institute at the Naval Postgraduate School, the University of North Carolina, Chapel Hill, and the Sarnoff Corp. (now part of SRI). The goals of BASE-IT are to address some of these difficulties and:

- to improve the preparation of trainees before their arrival,
- to supplement the information available to instructors, both in real-time during the exercise and for after-action review (AAR), and
- to automatically generate AAR resources for individual feedback.

The main components of BASE-IT include an automated camera management and sensing system, automated individual performance evaluation, automated analysis of unit behaviors, 3D visualization of recorded data sets with the ability to search for significant events, and automated behavior synthesis for exploration of 'what-if' scenarios.

Fig. 1. Observing in a "prone" posture, crossing a danger area, and results from our posture recognition method

This paper focuses on the extensive video analysis component, how it supplements the information available to instructors, and how it facilitates understanding of individual and team performances exhibited on the training range.

2 Related Work

Military training and performance measurement has long received tremendous attention. BASE-IT was built together with the Marine Air Ground Task Force Training Command at the Marine Corps Air Ground Combat Center at Twentynine Palms, California, which has one of the most advanced training facilities of the nation. Similar in instrumentation but without the analysis component is the Future Immersive Training Environment (FITE) at, Camp Pendleton's I Marine Expeditionary Force in California and at Camp Lejeune, NC (see, for example, [2, 3]). FITE's focus is on providing a training experience through augmentation, whereas BASE-IT as discussed here focuses on providing augmentations to the instructors and, particularly, to help analyze training. The US Army has similar training range instrumentations, for example, the Combat Training Center Military Operations on Urban Terrain Instrumentation System (CTC MOUT-IS) video system [4].

3 Solutions for Overcoming Cognition Limitations

Here we discuss the six limitations of unaided human cognition and what solutions we have applied to augment instructor cognition.

3.1 Omnipresence

The simulated town at the US Marine base at Twentynine Palms is nearly one square mile large and has many roads, houses, creeks, and other typical elements of any inhabited location. This prohibits instructors to have good visibility of all locations. We installed pole mounted fixed and PTZ cameras – a "sea of cameras" – to achieve omnipresence even in otherwise view-obstructed locations. Omnipresence through cameras provides augmentation of the spatial field of information. However, while this enables an instructor to virtually be at any one

of multiple locations, paying attention to multiple data streams simultaneously is difficult at best. Hence, we also require some degree of automated analysis of these additional data streams.

3.2 Continuous, Always-On Coverage

Human observation of dozens of live video feeds is impracticable if not infeasible due to the number of cameras and the continuous, always-on, non-stop 72 hours training scenario. Instead, we trained computer vision methods on the specific clothing, backpacks and helmets to detect US Marines and to estimate various body posture and weapon parameters, thereby filtering out empty scenes and scenes without any trainees. Night-time operations were observed to the degree possible with visible-spectrum cameras, plus the GPS and accelerometers on trainees and weapons. This presumably improves the instructor's ability to absorb information, essentially expanding the temporal horizon ("always-on"), the temporal resolution (several measurements per second), and the spatial extent (omnipresence). Despite the increased spatio-temporal field of view, information processing (see subsequent subsections) keeps the data volume manageable, and cognition augmented.

3.3 Posture Recognition and Head Localization

Fast action, multiple trainees, and the instructor's vantage point often prohibit precise estimates for the trainees' body and head positions. Yet these are important, for example, in order to determine whether a Marine has sought sufficient cover in case of enemy fire. We built posture recognition methods that can determine whether a Marine is standing or taking a knee, and we custom-trained head detection methods on the specific helmets to precisely locate them in the 3D environment [5, 6]. This offloads the spatial reconstruction task from the instructor to the computer and permits eyes-on *more* trainees at any time.

3.4 Monitoring Security

It is vital for US Marines to maintain "360° security" at all times, requiring coordination between individuals to visually scan in all directions as a team. Again, it is difficult if not impossible for instructors to assess this continuously, particularly if part of a team is hidden from view. Our computer vision methods automatically estimate the torso (shoulder) orientation and the head orientation of each trainee. A subsequent performance analysis module [7] monitors this information for the entire team and flags incidents of likely incomplete situational awareness. Cognition is augmented spatially again, around corners and through occlusions. It is also augmented through simultaneous assessment of head orientations of *all* squad members and automated calculation of the 360° coverage.

3.5 Identifying Weapon Flagging

One of the performance traits that is continuously observed and evaluated is flagging – unintentionally pointing a weapon system towards a fellow US Marine. Our system continuously determines the orientation of the weapon system with acceleration sensors and vision-based processing. It then checks against known positions of nearby US Marines, and identifies the times and places where incidents of flagging happened, including the identification of the individual who caused each flagging incident. Such a list of incidents speeds the instructor's comprehension of the trainee's performance. Further, it provides a second, unbiased look at trainees through the eyes of other modalities.

3.6 Foot Patrol Analysis

Another important team behavior concerns patrol formations and their dispersion across the terrain, that is, the distance between individual trainees and their spatial configuration. Depending on the situation, it is more or less dangerous to be close to each other or further apart, to walk in single file or offset, and so on. Similarly, foot patrols need to "cross danger areas" in a particular fashion: running, not walking, and not all at once (see Fig. 1). The BASE-IT performance analysis module utilizes the precise position estimates from our visual analysis to measure distances and velocities and to provide pre-analyzed results to the instructors. Again, these objective measurements supplement the subjective and often incomplete instructor's observations.

4 Results

Does BASE-IT indeed augment the instructors' cognition? This hypothesis can ultimately only be answered by directly measuring the cognition, either through objective means or through a questionnaire that assesses cognition. Neither of these options was viable for BASE-IT due to time and financial constraints, as well as due to the difficulty of constructing a control group of instructors for these one-time training actions. The approach taken here determines whether the instructors were given information that conceivably would *result* in augmented cognition.

Figure 2 depicts the various augmentations to the information available to instructors: additional spatial and temporal information, information in additional sensing modalities, and, last but not least, pre-processing and filtering of information. But let us take a closer look.

Spatial Augmentation. Merely providing information about previously inaccessible areas can suffice to make instructors cognizant of a situation they had no previous information on. For example, a foot patrol formation that was previously out of view comes into view with the help of our cameras. Provided the instructor looks at the imagery, he will become cognizant of this information. He will be able to have eyes on more trainees and avoid occlusions.

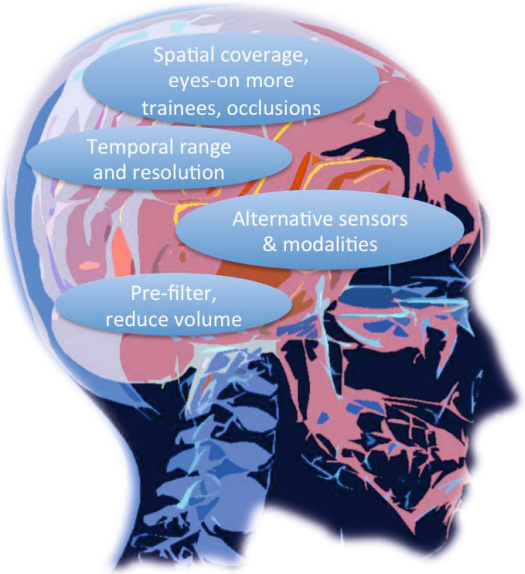


Fig. 2. Augmentations to human cognition

Temporal Augmentation. The cameras provide continuous coverage in lit areas, not taking a break and not getting tired. They also capture events at a frame rate that permits analysis of short-lived actions such as a brief weapon flagging event. As before, this increased temporal duration and resolution of data makes information accessible to the instructors, but they still have to actively seek out this information.

Sensory Augmentation. Since the trainees are tracked by visual information and GPS sensors, and their weapons' orientations are tracked with specific "Inertial Navigation System" (INS) sensors mounted on the weapons, "more than what meets the eye" helps determine locations and orientations of US Marines and weapons. These additional sensors and measurement modalities again increase the *amount* of information available to the instructors.

Pre-cognitive Filtering. Naturally, more information does not directly result in better understanding or cognition, just like the wealth of information available on the internet does not immediately translate into smarter surfers. However, filtering the information to only the most relevant aspects and thereby reducing its amount increases the chances that an instructor will find time to inspect it, especially if this information cannot easily be gleaned from other sources. Similarly, a site of distinct and mostly relevant information is more likely to be visited. BASE-IT provides pre-processed information that obviates the need for tedious video review and instead makes the most pertinent information available immediately. For example, presenting instances of weapon flagging or cases of "bunching up" is clearly much more useful than requiring review of hours of mostly uneventful video.

Note that the computer does not produce final results or even make the decisions. Instead, automated visual analysis and filtering "merely" improves the scene that is presented to the human. This is an important consideration for applied computer vision systems since it is unrealistic to expect perfect performance when translating recent research into practical application. Note also that BASE-IT distinguishes the two phases of data acquisition and processing in the terms "sensing" and "sense making." Sensing includes sensor management system, tracking of individuals (including pose and posture), and sense making includes automated behavior analysis and performance evaluation. Omnipresence and continuous coverage fall under the "sensing" aspect and the remaining solutions are mostly "sense making," albeit they use the additional sensor data from video or accelerometers, for example.

To illustrate the capability of pre-processing, we repeat here the results of the BASE-IT automated video analysis [6]. Using the automated video analysis, Marines were detected successfully (in uncluttered conditions) in 98.73% of the tested instances. When the subjects were partially occluded, the recognition was negatively impacted and only 53% of the torso orientations were correctly identified. The number of correctly classified instances (per marine and per frame) was determined to be 76% and 72% for the torso and head, respectively (see the confusion matrices in [6]). Speed performance tests showed that the detection task was accomplished in 1.9 seconds and that it scaled sub-logarithmically with an increase in image size. The combination of per-frame detection and posture recognition with semantic consistency checking and temporal smoothing [5] provides sufficient accuracy for determining tracks. These tracks can then be analyzed further for troop formation [7]. This is a task that is difficult to perform for human instructors, as discussed in Sec. 3.6, hence we consider BASE-IT augmenting the instructor's cognition.

By stressing salient activities and filtering out unimportant aspects we reached our objective of radically improving the control of and insight into the training exercise, enabling detailed after action review within minutes of completion of the exercise, and further enhancing and supplementing an already invaluable training experience.

5 Conclusions

Training US Marines for complex situations requires training in a complex environment, which poses a great challenge to instructors and their ability to assess the trainees accurately. In this paper, we described how BASE-IT attempts to improve upon the information available to instructors in the hope that it improves their understanding and analysis of the trainees.

Our experience shows that only in tandem does more information and its pre-processing truly augment cognition. BASE-IT pays specific, uninterrupted attention to individuals, anywhere on the range, with help of a multi-modal sensor suite, and through multi-stage analysis modules. BASE-IT provides value as a tool for both instructors and trainees, both for training preparation and for personalized review and analysis (AAR). In the near future, we expect many more

tools that pre-process the "big data" from training observations and, together with a human in the loop, permit semi-automatic analysis and much-improved feedback to the trainees.

Acknowledgements. This work was funded under ONR-BAA-05-023. We thank the volunteer contributors and excellent collaborators in the US Marine Corps without whom this project would not have been possible.

The views expressed in this document are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

[1] Sadagic, A., Kölsch, M., Welch, G., Basu, C., Darken, C., Wachs, J.P., Fuchs, H., Towles, H., Rowe, N., Frahm, J.M., Guan, L., Kumar, R., Cheng, H.: Smart Instrumented Training Ranges: Bringing Automated System Solutions to Support Critical Domain Needs. Journal of Defense Modeling and Simulation (2012) (accepted for publication)

[2] Livingston, M.A., Rosenblum, L.J., Brown, D.G., Schmidt, G.S., Julier, S.J., Baillot, Y., Swan II, J.E., Ai, Z., Maassel, P.: Military Applications of Augmented Reality. Springer, Heidelberg (2011)

[3] Muller, P., Schmorrow, D., Buscemi, T.: The Infantry Immersion Trainer – Today's Holodeck (September 2008)

[4] Blake, J.T.: Products and Services Catalog (2010), `http://www.peostri.army.mil/`

[5] Wachs, J.P., Kölsch, M., Goshorn, D.: Human Posture Detection for Intelligent Vehicles. Journal of Real-Time Image Processing (2010)

[6] Wachs, J.P., Goshorn, D., Kölsch, M.: Recognizing human postures and poses in monocular still images. In: Proc. Intl. Conf. on Image Processing, Computer Vision, and Signal Processing, IPCV 2009 (2009)

[7] Rowe, N., Houde, J., Kölsch, M., Darken, C., Heine, E., Sadagic, A., Basu, A., Han, F.: Automated assessment of physical-motion tasks for military integrative training. In: Second International Conference on Computer Supported Education, Valencia, Spain (2010)

A Novel HCI System Based on Real-Time fMRI Using Motor Imagery Interaction

Xiaofei Li[1], Lele Xu[1], Li Yao[1,2], and Xiaojie Zhao[1,*]

[1] College of Information Science and Technology, Beijing Normal University, Beijing, China
{frankdehao,xulelebnu}@gmail.com, zhaoxj86@hotmail.com
[2] State Key Laboratory of Cognitive Neuroscience and Learning,
Beijing Normal University, Beijing, China
yaoli@bnu.edu.cn

Abstract. Real-time functional resonance imaging (rtfMRI) provides an emerging human-computer interaction (HCI) technology with relatively high spatial resolution. The motor imagery is widely used for sports training of athletes and motor ability rehabilitation of patients, which is a common interaction approach for EEG-based and fMRI-based BCI. An appropriate method of interaction can improve the performance of BCI. In this paper, we implemented a novel HCI system based on rtfMRI using motor imagery interaction. The user interacted with the system by regulating blood oxygenation level dependent (BOLD) signal intensity of the region of interest (ROI) in motor areas using motor imagery, which was presented by the running speed of a virtual human in an animation. The ROI was chosen according to the motor network resulted from the real-time independent component analysis (rtICA). Through the interaction with the HCI system, the user could learn the effectiveness of his motor imagery.

Keywords: HCI system, real-time fMRI, motor imagery, animation interaction.

1 Introduction

The emerging intelligent human-computer interaction (HCI) technology based on cognitive neuroscience is a promising tool to provide more novel interactive experience [1, 2]. Brain-computer interfaces (BCI) based on Electroencephalography (EEG) have been used for volitional regulation of electrical brain activity [3, 4]. However, the regional specificity of self-regulation is limited to the relatively low spatial resolution of EEG. The BCI based on real-time functional resonance imaging (rtfMRI) is another non invasive BCI with comparatively high spatial resolution, which has been broadly used in the novel neuroscience investigations [5, 6] and potential clinical applications [7, 8].

Motor imagery [9] has been used for sports training of athletes and motor ability rehabilitation of patients, which is a common interaction approach in the EEG-based BCI [3, 4] and fMRI-based BCI [10-12]. As known in fMRI studies, motor imagery will lead to the activation of motor areas in brain, among which the motor areas such

* Corresponding author.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 703–708, 2013.

as primary motor cortex (M1) can be self-regulated through rtfMRI [5, 6]. Besides, an appropriate method of interaction can effectively improve the performance of BCI. Generally, the animation presentation can largely avoid the dull interactions (e.g. word, number) that are unattractive to the user.

In this paper, we implemented a novel HCI system based on rtfMRI using motor imagery interaction. The user interacted with the system by regulating blood oxygenation level dependent (BOLD) signal intensity of the region of interest (ROI) in motor areas using motor imagery, which was presented by means of the running speed of a virtual human in an animation. The ROI was chosen according to the motor network resulted from the real-time independent component analysis (rtICA) [13, 14]. Through interaction with the HCI system, the user could learn the effectiveness of his motor imagery and try different strategies to adjust his brain.

2 Methods

2.1 The HCI System

The HCI system based on rtfMRI consists of hardware and software (Fig. 1). The MRI scanner scans the user's whole brain every repetition time (TR). Within one TR, the software has to finish the data preprocess and statistical analysis, and control the interactive presentation in the projector displayed to the user. The data preprocess includes head motion correction and spatial smooth, which is to reduce the noise in signal. The method of statistical analysis is sliding-window rtICA [14], which can result in the task-related brain network. The interactive control is to compute the running speed in the animation of a virtual human running (Fig. 2) in accordance with the BOLD signal originating from the corresponding motor areas.

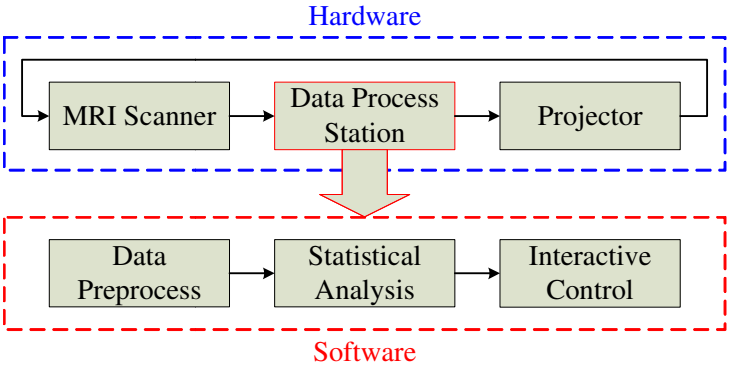


Fig. 1. The HCI system based on rtfMRI

Fig. 2. The 12 frames of pictures in the running animation

2.2 The Experiment Design

We carried out an experiment in block design on one subject (female, age 21 years), which was performed in a 3.0-T Siemens MRI scanner. The 32 axial slices were acquired in an interleaved order using a single-shot T2*-weighted gradient-echo EPI (echo-planar imaging) sequence (TR/TE/flip angle = 2000ms/40ms/90°, matrix size = 64×64, voxel size = 3.1×3.1×4.8 mm^3, slice thickness = 4 mm, slice gap = 0.8 mm).

The experiment included three sessions. The first session was for ROI functional localization, in which the subject was instructed to tap his right hand fingers. The following two sessions were for animation interaction, in which the subject regulated the running speed using motor imagery. Here, we took the first-person motor imagery that included both visual and kinesthetic imagery compared with the third-person motor imagery [15, 16].

In the ROI functional localization session, a motor network was resulted from the rtICA, in which the supplemental motor area (SMA) was chosen as the ROI for self-regulation. The entire session was made up of five 30s rest blocks and four 30s task blocks. Each block consisted of fifteen trials and each trial lasted 2s. During the rest blocks, a text cue "rest" was presented to the subject, and during the task blocks, a text cue "task" was presented.

The interaction session was made up of eight 30s rest blocks and seven 30s task blocks. Each block consisted of fifteen trials and each trial lasted 2s. During the rest blocks, a green cross was presented in the center of the monitor and the subject was instructed to take a rest and think nothing. During the task blocks, the running interactive animation was presented and the subject was requested to adjust the speed as fast as possible. It was allowed that the subject could try different motor imagery strategies (e.g., playing basketball, playing piano) to reach a high running speed.

2.3 The Interactive Control

In the animation, one complete virtual human running action is made up of twelve frames of pictures (Fig. 2). The animation is performed by changing the frame rate (FPS) with the ROI activation intensity (S). The running speed is termed by the number (N) of complete actions in one second. The formulation of FPS and N can be described as follows:

$$FPS = N \times 12 \tag{1}$$

The average signal intensity of the ROI in the last rest block is taken as a baseline (B). The ROI activation intensity (S) is the signal intensity in one scan of the current task block, and the signal change (C) is $S - B$. Thus, the N can be determined as follows:

$$N = N_{baseline} + \frac{0.5N_{\max}}{C_{\max}} \times C \tag{2}$$

In practice, the maximum $N(N_{\max})$ is chosen as two to prevent the running speed from getting too fast for the user. The baseline of $N(N_{baseline})$ is chosen as one when

C is zero, and the bottom of N is zero. Since the signal change in real motion is generally stronger than that in motor imagery, we choose the maximum C in the first localization session as the $C_{\max}$ for the follow interaction sessions.

3 Results

In the ROI functional localization session, the motor network was automatically derived from the rtICA, which mainly included the left M1, pre-motor cortex and SMA (Fig. 3). The SMA was chosen as the ROI for the next two interaction sessions.

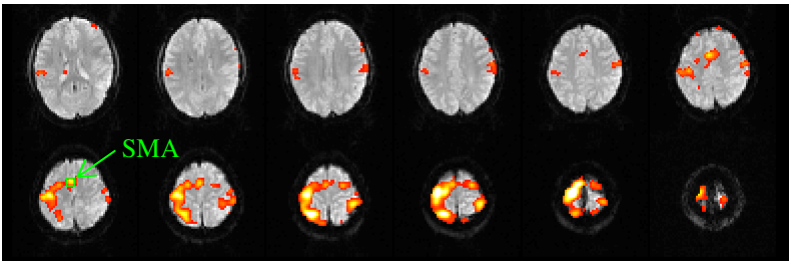


Fig. 3. The brain network derived from the sliding-window rtICA. The green rectangle was the chosen ROI (SMA).

In the second interaction session, the running speed of the animation was computed from the signal change of the target ROI by the formulations above (Fig. 4). In most of the task blocks, it could be seen that the signal intensity arose at the beginning of the block, and then turned fluctuant. The running speed changed with the signal intensity, and was faster than baseline as a whole in majority of the task blocks.

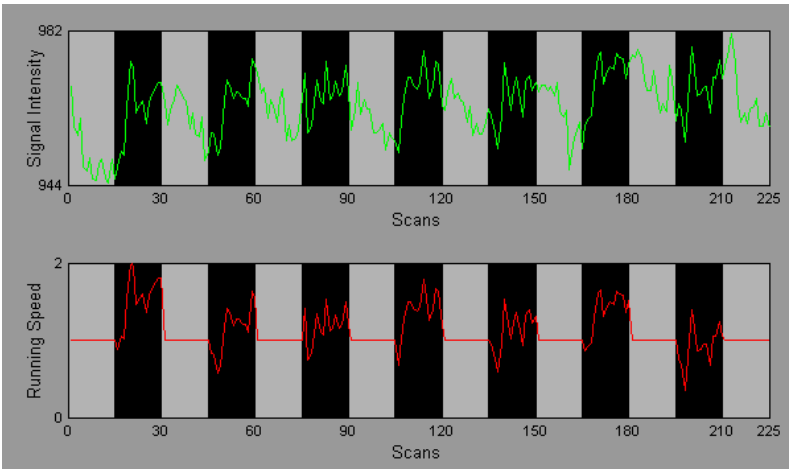


Fig. 4. The average signal intensity change of ROI (up, green curve) and the running speed change of the animation (down, red curve) along with the scans in the second interaction session. The black blocks were task blocks and the gray ones were rest blocks. Besides, the baseline of running speed is one as shown in the straight line in the rest blocks (down).

4 Discussion

We successfully applied the HCI system in an interactive motor imagery experiment. After the experiment, the subject reported that the running speed could be basically controlled by his motor imagery at the task blocks, which was in accord with the Fig. 4. It was hard for the subject to concentrate on the task all the time, so the speed would fall down or be out of control at times when he was distracted. Nevertheless, the subject still expressed that the animation was very interesting and novel. It could be seen that the experience of observing and controlling one's own brain activity might bring more fun in the interactive process.

As a new way to BCI technology, rtfMRI-based motor imagery could be used to control the movement of a cursor turning left or right through a maze [11] and the movement of a robotic arm [12]. In our system, the running speed depending on the intensity of SMA could also serve as a control signal to manipulate real machines. This showed the feasibility to extend this online system.

Compared with the previous studies on rtfMRI which mainly focused on the ROI activated by the task, our work also provided a new way to assess and control the activation of the task related network. In the component derived from the ICA (Fig. 3), as an important node in motor network [17], SMA was chosen as the target ROI to be regulated as well as the regulation of motor brain network. The whole component could also be chosen as the target to be regulated. Thus, the system using ICA method might have a potential for interaction based on brain network activities.

After all, the experiment was performed on only one subject, so the result was very preliminary. The effectiveness of motor imagery interaction has to be further investigated in a larger sample size with more sessions to allow for the conclusions on the efficiency of this HCI system.

Acknowledgment. This work was supported by Key Programs of the Nature Science Foundation of China (NSFC) with Project Number 60931003, and the General Program of NSFC 61071178.

References

1. Lebedev, M.A., Nicolelis, M.A.L.: Brain-machine interfaces: past, present and future. TRENDS in Neurosciences 29, 536–546 (2006)
2. Graimann, B., Allison, B., Pfurtscheller, G.: Brain-computer interfaces: Revolutionizing human-computer interaction. Springer (2011)
3. Pfurtscheller, G., Neuper, C.: Motor imagery and direct brain-computer communication. Proceedings of the IEEE 89, 1123–1134 (2001)
4. Christa Neuper, Ã., Muller-Putz, G.R., Scherer, R., Pfurtscheller, G.: Motor imagery and EEG-based control of spelling devices and neuroprostheses. Event-Related Dynamics of Brain Oscillations 159, 393 (2006)
5. Yoo, S.S., Jolesz, F.A.: Functional MRI for neurofeedback: feasibility study on a hand motor task. Neuroreport 13, 1377–1381 (2002)

6. Decharms, R.C., Christoff, K., Glover, G.H., Pauly, J.M., Whitfield, S., Gabrieli, J.D.E.: Learned regulation of spatially localized brain activation using real-time fMRI. NeuroImage 21, 436–443 (2004)
7. DeCharms, R.C., Maeda, F., Glover, G.H., Ludlow, D., Pauly, J.M., Soneji, D., Gabrieli, J.D.E., Mackey, S.C.: Control over brain activation and pain learned by using real-time functional MRI. Proceedings of the National Academy of Sciences of the United States of America 102, 18626 (2005)
8. Haller, S., Birbaumer, N., Veit, R.: Real-time fMRI feedback training improve chronic tinnitus. European Radiology 20, 696–703 (2010)
9. Decety, J.: The neurophysiological basis of motor imagery. Behavioural Brain Research 77, 45–52 (1996)
10. Weiskopf, N., Mathiak, K., Bock, S.W., Scharnowski, F., Veit, R., Grodd, W., Goebel, R., Birbaumer, N.: Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI). IEEE Transactions on Biomedical Engineering 51, 966–970 (2004)
11. Yoo, S.S., Fairneny, T., Chen, N.K., Choo, S.E., Panych, L.P., Park, H.W., Lee, S.Y., Jolesz, F.A.: Brain-computer interface using fMRI: spatial navigation by thoughts. Neuroreport 15, 1591 (2004)
12. Lee, J.H., Ryu, J., Jolesz, F.A., Cho, Z.H., Yoo, S.S.: Brain-machine interface via real-time fMRI: preliminary study on thought-controlled robotic arm. Neuroscience Letters 450, 1–6 (2009)
13. Esposito, F., Seifritz, E., Formisano, E., Morrone, R., Scarabino, T., Tedeschi, G., Cirillo, S., Goebel, R., Di Salle, F.: Real-time independent component analysis of fMRI time-series. NeuroImage 20, 2209–2224 (2003)
14. Ma, X., Zhang, H., Zhao, X., Yao, L., Long, Z.: Semi-Blind Independent Component Analysis of fMRI Based on Real-Time fMRI System. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 1 (2012)
15. Fourkas, A.D., Avenanti, A., Urgesi, C., Aglioti, S.M.: Corticospinal facilitation during first and third person imagery. Experimental Brain Research 168, 143–151 (2006)
16. Sirigu, A., Duhamel, J.R.: Motor and visual imagery as two complementary but neurally dissociable mental processes. Journal of Cognitive Neuroscience 13, 910–919 (2001)
17. Sharma, N., Baron, J.C., Rowe, J.B.: Motor imagery after stroke: relating outcome to motor network connectivity. Annals of Neurology 66, 604–616 (2009)

Guided Learning Algorithms: An Application of Constrained Spectral Partitioning to Functional Magnetic Resonance Imaging (fMRI)

Henry L. Phillips[1], Peter B. Walker[1], Carrie H. Kennedy[2], Owen Carmichael[3], and Ian N. Davidson[3]

[1] United States Navy
[2] Marine Corps Embassy Security Group
[3] University of California – Davis
{henry.phillips,peter.b.walker}@navy.mil, carrie.kennedy@usmc.mil, ocarmichael@ucdavis.edu, davidson@cs.ucdavis.edu

Abstract. Innovations in neuro-technology have created a potential gap in our ability to measure human performance and decision making in dynamic environments. Therefore, a need exists to create more reliable testing methodologies and data analytic solutions. The primary aim of this paper is to describe work to integrate subject matter expertise with algorithms designed to measure human brain activity in real time. Specifically, Guided Learning using constrained spectral partitioning to increase the reliability and interpretability of fMRI data is explicated and applied as a test case to the Default Mode Network in the elderly population. How Guided Learning can be further applied to other neuroimaging technologies that may be more conducing to furthering the field of augmented cognition is discussed.

Keywords: augmented cognition, functional connectivity, fMRI.

1 Introduction

Since its inception as a scientific field at the turn of this century, augmented cognition has been one of the fastest growing research areas influencing several different academic disciplines including engineering, psychology, and human factors [1]. The excitement surrounding this field of research has allowed for an explosion of innovation in the ability to capture human performance and decision making through innovations in neuro-technology. However, a gap may soon develop as researchers attempt to develop research methodologies to integrate these innovations into the laboratory environment. The primary aim of this paper is to describe progress integrating subject matter expertise (SME) with algorithms designed to measure human brain activity in real time. We term this work Guided Learning to reflect the pursuit to develop a general class of algorithms that incorporate SMEs to help identify meaningful and insightful patterns within dense datasets.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 709–716, 2013.

For the research described in this paper, we will focus on the analysis and use of functional Magnetic Resonance Imaging (fMRI). fMRI provides the unique opportunity to visualize neural activity in the brain in an on-line modality. Perhaps one of the most promising applications of fMRI data has been on the analysis of functional connectivity [2], the temporal correlation of neuronal activation for spatially discrete locations. The excitement over this analytic approach is due, in part, to the applicability of these findings in both clinical and diagnostic settings. For example, functional connectivity studies have explored issues such as the study of post-traumatic stress disorder (PTSD), chronic traumatic encephalopathy (CTE) and traumatic brain injury (TBI) [3].

However, a review of the literature on the use of fMRI in the analysis and interpretation of functional connectivity data illustrates several limitations in the current methodology used to interpret this data. For example, analysis of functional connectivity data is often limited by poor test/re-test reliability. As is seen in Figure 1, results from separate scans may yield different results. In this figure, we see the results for two resting state fMRI data sets of the same healthy young individual, acquired in short succession. Barring a major medical event between the two scans, the spatial and temporal patterns of resting state activity should be very similar. Yet, the data analytic approach used on this data set identified different patterns of activation across the two sets of scans.

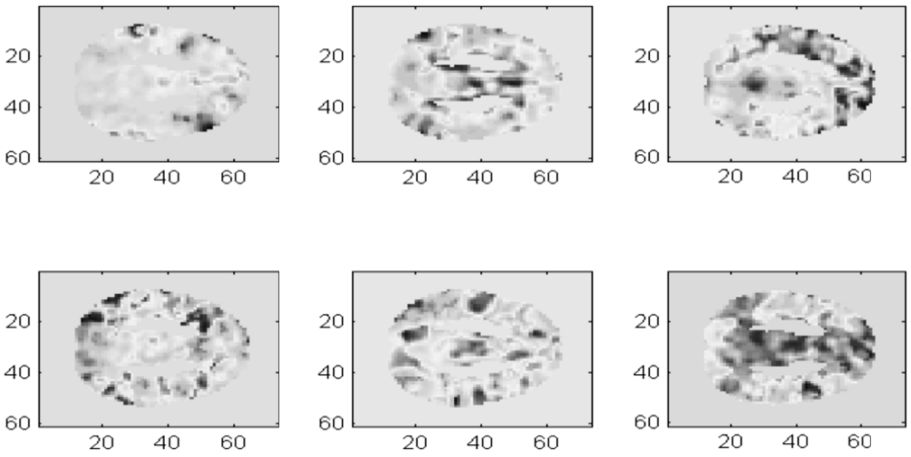


Fig. 1. Lack of test-retest reliability in fMRI. A healthy young individual received two fMRI scans in rapid succession. The top and bottom rows show the three top-ranked clusterings for the first and second scans.

Most often, datasets such as those illustrated in the figure above are modeled as a graph: each node corresponds to an instance and (the weight of) each edge corresponds to the similarity between two instances. To partition the dataset, a commonly used algorithm is spectral clustering [4], which finds the (relaxed) normalized min-cut of the corresponding graph. Traditional spectral clustering only applies to a single graph. However, in a wide range of applications, in addition to the graph under

consideration, there is auxiliary information available in the form of a second graph, which shares the same set of nodes with the first graph, but has a different set of edges. A number of alternatives exist under which a second graph might meet these assumptions, including 1) the edges of the second graph are constructed based on a different set of features; 2) the edge weights of the second graph are computed using different similarity functions; and/or 3) the two graphs represent the evolution of a graph over time. Intuitively, the extra knowledge from a second graph may help to identify a better partition than the best one that can be identified using the first graph only [5]. As will be explained later, for optimal utilization, this information must be partially theoretically driven, using qualitative input from a SME, rather than derived solely through algorithm application.

A direction already explored by the community is to consider any two such graphs as two independent views and combine them into one graph, to which the traditional spectral clustering algorithm is then applied [6]. However, this approach relies on the assumption that the two views are complementary and thus helpful to each other, which is not always the case in practice. The approach outlined in this paper attempts to transfer "knowledge" from more stable fMRI images to those scans where a particular activation function may not be as readily apparent. To accomplish this, a form of spectral clustering to two separate groups of fMRI scans was applied. Unlike traditional clustering algorithms, such as spectral clustering that attempt to segment a graph based on a single image, our approach incorporates knowledge from multiple graphs that might share the same set of nodes with the first graph, but have a different set of edges.

2 Limitations of Spectral Clustering

Clustering analytic approaches remain one of the most widely used techniques for exploratory data analysis and have been used extensively in areas ranging from image processing to functional connectivity analysis [7,8,9]. Furthermore, some forms of clustering may be more applicable to particular problem sets over others. For example, spectral clustering has been argued to be superior to traditional clustering algorithms like K-means because it yields a deterministic polynomial-time solution, provides researchers the ability to model arbitrary shaped clusters, and affords equivalence to certain graph cut problems.

However, as mentioned previously, traditional clustering like spectral clustering can only be applied to a single graph. In a wide range of applications, such as the analysis of several distinct fMRI scans, it would be more beneficial to combine properties from different graphs to form a single cut from the data comprising the set. This approach, which has only been recently introduced by the clustering community, has come to be known as constrained spectral clustering.

Constrained spectral clustering attempts to incorporate auxiliary information from separate graphs to help improve clustering on both graphs. In general, constrained clustering is a category of techniques that tries to incorporate Must-Link (ML) and Cannot-Link (CL) constraints into existing clustering algorithms. It has been well

studied on algorithms such as K-means clustering, mixture modeling, hierarchical clustering and density-based clustering.

Multi-view Spectral Clustering

In contrast to constrained spectral clustering, traditional multi-view spectral clustering algorithms attempt to consider a set of two graphs as two independent views and combine information from both graphs into one graph. However, in its basic form, this relies on the assumption that the two graphs are complementary to one another. That is, it is assumed that both graphs are noise-free. In the work presented in this paper, we no longer assume the two graphs are complimentary. Rather, our approach, which we term *Constrained Spectral Partitioning* (hereafter, CSP) attempts to discover an alternative direction of finding a cut whose edge weight is minimized based on information about both graphs [10, 11, 12].

We contend that CSP fits into a general category of algorithms, i.e., Guided Learning. This term is appropriate because, in addition to algorithm application, we allow for SME input to maximize the identification of appropriate cut(s) for a series of scans. Assume we have two graphs, an exemplar graph and a target graph, that share the same set of nodes but have different sets of edges or edge weights. The goal of applying SME input, is the utilization of information from the exemplar graph to identify a more representative and replicable cut on the target graph.

Further, we believe this work represents a hitherto unattempted technique to "close the loop" with respect to Augmented Cognition. Previous work in algorithm development for Augmented Cognition was focused on utilizing machine learning to maximize human performance. However, our work here attempts to close the loop by allowing for SME input to further maximize the efficiency of the learning algorithm.

3 Background and Graph Theory Notation

Formally, the set of points in a network may be represented as a weighted undirected graph $G = (\boldsymbol{V}, \boldsymbol{E})$, where the nodes are the set of points in a feature space and an edge is formed between each pair of nodes. The weight (similarity) on each edge $w(\boldsymbol{i}, \boldsymbol{j})$ is a function of the similarity between nodes $\boldsymbol{i}$ and $\mathbf{j}$.

To more effectively interpret the graph, grouping or clustering techniques may be applied that attempt to segment the graph into more similar sub graphs containing similar features. This may be accomplished by partitioning the graph into multiple disparate sets $\boldsymbol{V}_1, \boldsymbol{V}_2,\ldots,\boldsymbol{V}_m$, where some measure of similarity is high among vertices within set $\boldsymbol{V}_i$ but very low across different sets of vertices between sets $\boldsymbol{V}_i$ and $\boldsymbol{V}_j$.

However, as was discussed previously, the traditional approaches do not take into account those cases where we may only wish to extract certain features from some of the graphs. In this section, we describe how we adapted the classical spectral clustering to increase reliability when segmenting across one or several graphs.

To accomplish this, CSP was applied such that one or several source graphs were identified and used to segment several target graphs. The knowledge to transfer was derived from the source graph in the form of what we termed, degree-of-belief

constraints. Specifically, let $G_S(\boldsymbol{V}, \boldsymbol{E_S})$ be the source graph and $G_T(\boldsymbol{V}, \boldsymbol{E_T})$ the target graph. A_S and A_T are their respective affinity matrices. Then, A_S can be considered a constraint matrix with only *ML* constraints. It carries the complete knowledge from the source graph, and we can transfer it to the target graph using our constrained spectral clustering formulation:

$$\underset{\mathbf{v}\in\mathbb{R}}{argmin}\, \mathrm{v}^T \bar{L}_T\mathbf{v}, s.t.\, \mathbf{v}^{\mathrm{T}}\mathrm{A_s}\mathbf{v} \geq \alpha, \mathbf{v}^{\mathrm{T}}\mathbf{v} = \mathrm{vol(G)}, \mathbf{v} \neq \mathrm{D_T^{1/2}}\mathbf{1}$$

α is now the lower bound of how well the knowledge from the source graph must be enforced on the target graph. The solution to this is similar:

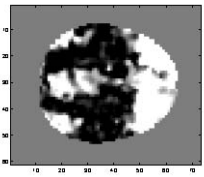
$$\bar{L}_T\mathbf{v} = \lambda(\bar{\mathrm{A}}_{\mathrm{S}} - \frac{\beta}{\mathrm{vol}(\mathrm{G_T})}\mathrm{I})\mathbf{v}$$

Note that since the largest eigenvalue of $\bar{\mathrm{A}}_{\mathrm{S}}$ corresponds to a trivial cut, in practice we should set the threshold such that $\beta < \lambda_1 \mathrm{vol(G)}$, $\lambda 1$ is the second largest eigenvalue of $\bar{\mathrm{A}}_{\mathrm{S}}$. This will guarantee a feasible eigenvector that is not the trivial cut.

4 Application of CSP to fMRI Analyses

In this paper, we apply CSP to the analysis of the Default Mode Network [13]. The DMN is an interconnected brain system that activates simultaneously and periodically while in rest state. It has been hypothesized that the DMN is only active when individuals are focused on internal tasks such as daydreaming, memory retrieval, or introspection. The DMN is composed of several subsystems including part of the medial temporal lobe, medial prefrontal cortex, the posterior cingulated cortex, and the lateral and inferior parietal cortex.

The DMN provides a relevant test-bed for measuring the reliability/stability of the technique presented herein, as it has been reliably shown that the DMN is more pronounced and observable in young healthy patients than individuals suffering from various mental pathologies [14]. In this case, we compared DMN signatures across fMRI scans for young healthy patients with those suffering from early and late Alzheimer's disease. If CSP is able to reduce the impact of noise within fMRI scans across different abnormal groups, then it is hypothesized that CSP might increase the sensitivity to allow for the detection of specific phenomena – in this case, degree of similarity to an exemplar DMN among a set of Alzheimer's patients.



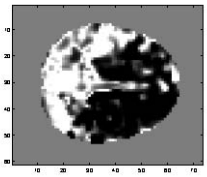


Fig. 2a and 2b. Figure 2a (left) displays segmentation results from a normal healthy participant. Figure 2b (right) displays segmentation results from a participant diagnosed with Alzheimer's disease. DMN activation appears as lighter colored pixels.

As shown in Figure 2a above, segmentation of a graph from a normal participant (P1) captures the DMN (the light pixels). However, if we apply spectral clustering to another graph constructed from an Alzheimer's patient's (P2) fMRI scan, the normalized mincut shows an entirely different pattern (Figure 2b).

Here, CSP was applied as a new approach for assessing inter-individual clustering commonalities at a population level. The principal benefit yielded by the reliance of an exemplar scan as the basis for partitioning decisions among target group members is an improvement in the reliability of intra-individual fMRI clustering in the target group. CSP incorporates user-provided guidance about which voxels should and should not cluster together. Our approach is to use the clustering of an exemplar scan to generate guidance (constraints), and use them in CSP to cluster a target scan. The exemplar scan is explicitly assumed to exhibit desirable or representative clustering behavior. If multiple diverse clusterings of the target scan all yield similar cut costs, CSP identifies the one most similar to that of the exemplar at each timepoint, yielding improved intra-individual clustering reliability across different scans.

We used real resting-state fMRI scans of young and elderly (Normal, Mild Cognitive Impairment, and Demented) individuals to demonstrate the advantages of CSP over spectral partitioning. In comparing the two groups, we applied segmentation algorithms so as to identify the DMN for each group of scans. As has been previously shown, we expected that the tightness of this clustering would be decreased in elderly individuals, especially those with Alzheimer's disease. Therefore, we identified an exemplar scan of a young individual whose spectral partitioning clearly indicated the DMN as one of its clusters. We then applied CSP to partition target scans including young and elderly individuals based on constraints derived from this exemplar.

In order to assess whether CSP increased reliability over and above the use of spectral partitioning alone, we first compared the test-retest reliability of the spectral partitioning with that of CSP on a group of individuals who received a pair of fMRI scans at two different time intervals. For each pair of fMRI scans, we calculated the percent difference in spectral partitioning and CSP costs between scans (i.e., the absolute difference in partition costs divided by average partition cost). The data from this study supported the claim that CSP increases reliability over and above what is found from spectral partitioning alone.

Our next analysis focused on assessing the biological validity of CSP. To assess the biological validity of CSP, we compared partition costs for each of three groups of participants: Elderly, Mild Cognitive Impairment, and Demented. Figure 3 below plots the measure of reliability within different groups of participants. As can be seen from the figure below, the average CSP cut cost was greater in MCI compared to healthy elders, and greater in dementia compared to MCI. The MCI partition costs were more variable, spanning most of the range of normal and demented values. The difference between normal elderly and demented cut costs was statistically significant ($p = .046$). This finding is consistent with previous findings that suggest that the DMN is less pronounced (therefore exhibiting higher cut costs) in individuals with Alzheimer's disease.

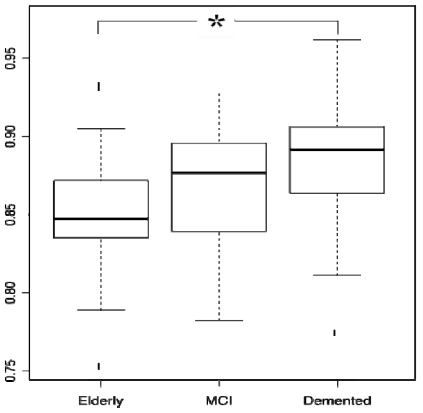


Fig. 3. Partition costs within healthy elderly, MCI, and demented groups. * Significant group difference at the $p<.05$ level.

Several key findings were presented above. First, we showed that deriving clustering constraints from an exemplar fMRI scan, and using them to cluster a target fMRI scan via CSP leads to resting state fMRI clustering results that are more reliable than traditional clustering approaches. In addition, we showed that the application of CSP in this dataset resulted in better identification of known biological differences between elderly individuals that are associated with neurodegenerative disease.

5 Guided Learning and Augmented Cognition

As newer technologies allow for the visualization of the brain during performance and decision making tasks, it is imperative that researchers incorporate paradigms that allow for more accurate assessments of network activation during these tasks. In this paper, we argued that the way forward for Augmented Cognition in this endeavor is a "closed loop" whereby SME input is incorporated with machine learning algorithms to measure and assess human performance. Algorithms such as CSP will provide augmented cognition researchers the opportunity to incorporate real and practical theory with basic science in the hopes of closing a potential gap between technology and the paradigms augmented cognition practitioners incorporate in their research.

Acknowledgements. The research reported in this paper was supported by Office of Naval Research Grants NAVY 00014-09-1-0712 and NAVY 00014-11-1-0108. The opinions of the authors do not necessarily reflect those of the United States Navy. We would also like to acknowledge Owen Carmichael from the Alzheimer's Disease Institute at the University of California – Davis for his assistance in data collection.

References

1. Schmorrow, D.D., Reeves, L.M. (eds.): HCII 2007 and FAC 2007. LNCS (LNAI), vol. 4565. Springer, Heidelberg (2007)
2. Friston, K.J., Frith, C.D., Liddle, P.F., Frackowiak, R.S.J.: "Functional Connectivity: The Principle Components Analysis of Large (PET) Data Sets. Journal of Cerebral Blood Flow and Metabolism 13, 5–14 (1993)
3. Hoge, C.W., Castro, C.A., Messer, S.C., McGurk, D., Cotting, D.I., Koffman, R.L.: Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. N. Engl. J. Med. 351(1), 13–22 (2004)
4. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
5. Wang, X., Davidson, I.N.: Active Spectral Clustering. In: Proceedings of the IEEE International Conference on Data Mining, ICDM 2010 (2010)
6. Bach, F.R., Jordan, M.I.: Learning spectral clustering. In: NIPS (2003)
7. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 1601–1608. MIT Press, Cambridge (2005)
8. Belliveau, J.W., Kennedy, D.N., McKinstry, R.C., Buchbinder, B.R., Weisskoff, R.M., Cohen, M.S., Vevea, J.M., Brady, T.J., Rosen, B.R.: Functional mapping of the human visual cortex by magnetic resonance imaging. Science 254, 716–719 (2006)
9. Stam, C.J., Reijneveld, J.C.: Graph Theoretical Analysis of Complex Networks in the Brain. Nonlinear Biomedical Physics 1, 3 (2007)
10. Basu, S., Davidson, I., Wagstaff, K. (eds.): Constrained Clustering: Advances in Algorithms, Theory and Applications. Data Mining and Knowledge Discovery, vol. 3. Chapman & Hall/CRC.Morup (2008)
11. Wang, X., Davidson, I.N.: Flexible Constrained Spectral Partitioning. In: KDD, pp. 563–572 (2010)
12. Wang, X., Qian, B., Davidson, I.: On Constrained Spectral Clustering and Its Applications. Journal of Knowledge Discovery and Data Mining (in press)
13. Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L.: The brain's default network. Annals of the New York Academy of Sciences 1124(1), 1–38 (2008)
14. Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V.: Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proceedings from the National Academy of Sciences U.S.A. 101, 4637–4642

Next Generation of Physical Training Environments: Bringing in Sensor Systems and Virtual Reality Technologies

Amela Sadagic

Naval Postgraduate School, 700 Dyer Road, 93943 Monterey, CA, USA
a.sadagic@nps.edu

Abstract. Training on physical training ranges is immensely important to any military unit, as many aspects of individual and team skills still need to be trained there. Nevertheless, the overall cost of training on physical ranges, the required unit throughput, as well as a need to maximize the training potential that such precious environments have, are ever increasing, and leveraging emerging technologies to make the training more effective becomes a necessity. This paper reviews several novel efforts in the research domain that could be used as a guide to the types of emerging technical solutions that may be employed to augment the capabilities of physical training ranges, including the capabilities of humans engaged in orchestrating and executing the training events (range operators and instructors), and to support the global objective of acquiring more effective training solutions.

Keywords: physical training ranges, sensor systems, virtual reality (VR), augmented reality (AR), automated behavior analysis, performance evaluation.

1 Introduction

Training on physical training ranges is immensely important to any military unit. These environments enable full skill integration; where elements of perceptual, cognitive and motor skills in the applied domain, and with a given objective, get to be practiced in situations that include physical exertion as well as real environmental conditions (sun, wind, temperature, humidity). Additionally, a unit's communications, cohesion and coordination, as well as its ability to operate and coordinate actions with other units gets decisively tested. By their nature, the physical training ranges are closest to the operational environments that are the final and ultimate goal for any unit. However, while training on physical ranges is an important form of training, it is often the most expensive form of unit preparation for future operations. It includes considerable material, logistical, as well as human resources, with zero to very minimal ability to reuse the same resources. Additionally, some elements of this type of training do not provide the flexibility that exists in computer-supported training interventions, thus reducing their overall effectiveness. It is therefore understandable that there is added pressure to maximize the training potential that such environments

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 717–726, 2013.

have, and to ensure that the most effective training does happen. One way of doing this is to employ different technical solutions, each addressing specific critical shortcomings of physical environments. Leveraging emerging technologies to enhance training capabilities is also identified as a one of the important aspects in supporting the central strategic idea of the Department of Defense (DoD) training concept for the next generation of DoD training solutions [1].

This paper reviews several novel efforts in the research domain that could be used as guidance on what types of emerging technology solutions may be employed to augment the capabilities of physical training ranges, including the augmentation of capabilities of humans engaged in orchestrating and executing the training sessions (range operators and instructors/evaluators).

2 Contemporary Training Solutions

Training solutions available to the training community can be divided into three general categories: (1) physical training environments (training ranges), (2) computer-supported training solutions (example: fully virtual simulations), and (3) a mix of both (example: physical environments augmented by synthetic elements, typically the visuals and sensor data feeds like those in bridge trainers). This section reviews current practices in two domains, one being a physical (physical training ranges), and another one being a virtual domain (computer-supported training solutions).

2.1 Physical Training Ranges

Out team conducted extensive observations of training courses organized on US Marine Corps physical training ranges, including the ranges for combined arms and urban warfare training [2]. As an example of physical training ranges that could greatly benefit from integrating emerging technologies in their daily operation, we will comment on ranges for rural and urban warfare; very similar conclusions could be drawn for training ranges for combined arms as well.

There are several characteristics associated with the space and training events organized on ranges for rural and urban warfare: (a) they are potentially large with a number of stationary objects, like buildings and urban and battlefield clutter, spread across them that cause a high level of occlusion, (b) a number of individuals (unit members in full combat gear, role players) and vehicles move around, contributing to increased complexity of training events, (c) some events can be staged (role players acting in certain areas), however the unit may decide to take a different route or take more time on a particular segment of their operations and may never experience the pre-planned interactions, and (d) while they do employ sensor systems, a predominant mode of operation is that the sensor management is human-driven (Figure 1 represents an illustration of a contemporary optical sensor system with multiple point-tilt-zoom (PTZ) and fixed cameras). The very nature of those environments is that they are highly occluded, and as a result some events can go unnoticed by instructors. In order to be able to keep good track of where everyone is and what everyone is

doing, a number of instructors (evaluators) are needed to conduct a single training session. In the case of long training courses (e.g. 72 hours), the instructors change, which inevitably brings a discontinuity in observation of unit performance.

Fig. 1. A segment of a camera system with one PTZ and three fixed cameras (*left*), one screen with video feeds from 12 cameras (*middle*), and a control room with an operator (*right*)

2.2 Computer Supported Training Solutions

Computer-supported training solutions, while not having many of the characteristics of physical training ranges, do bring elements that physical spaces do not have; providing opportunities for practicing a number of scenarios in a shorter period of time, saving precious material, logistical and human resources, as well as enabling training situations that would not be possible otherwise (practicing emergency procedures in a flight simulator is a good example), constitute some of those advantages. By the nature of this type of solution, all data in the system are already digital. This means that there are numerical values associated with all simulated and measured phenomena, which makes the tracking, analysis and queries of the individual or compound values much easier. Additionally, a standard that is already well established for these types of training solutions is that they are expected to integrate the elements of a training management system: they track the progress of each trainee in each session and across multiple sessions. All of this would be very hard, if not impossible, in physical training ranges.

Table 1 provides a basic comparison between physical training ranges and computer-supported training solutions. The differences identified in each aspect typically serve as the main resources and motivation for employing different emerging technologies most capable of filling those gaps. Our extensive observations of training courses on physical training ranges, helped us identify a number of issues that could be addressed by inserting emerging technologies, and our suggestions are based on data sets collected during those observations [2]. An added limitation on what type of technologies can be brought in to augment current training practices has a very hard requirement – no system or tool should get in the way of a unit's training objectives. If it does, then it is an unsuitable solution for that particular training situation. A good example of such solutions would be a wearable sensor that prevents a person from moving freely (running, crawling, kneeling, etc), something that is essential for the operations practiced on physical training ranges.

Table 1. Physical training ranges versus computer-supported training solutions

Aspect	Physical training ranges	Computer supported training solutions
Visibility of trainees' actions	Limited in occluded environments	All actions 'visible' to the system
Smart instrumentation	Minimal and operated by humans	Enabled (system tracks and initiates events)
Type of sensory data (visual, auditory, etc)	All analog, unless sensors used to detect events	All digital. Maintained and analyzed by the system
Performance evaluation	Instructor-driven only	Mix of automated (system-driven) and instructor-driven
Closed loop feedback	Instructor-driven only	Mix of automated (system-driven) and instructor-driven
After Action Review (AAR)	Instructors' commentary (what they wrote or remembered from training session)	System account of group and individual performances, combined with instructors' commentary
Take-away package	Instructors' notes and recorded video footage (multiple DVDs)	System recordings and system account of group and individual performances (digital, searchable data sets).
Training management system (tracking trainees' performance across multiple sessions)	Manual, basic information only	Automated (system driven) tracking and analysis possible
Historical trends analysis	Manual	Automated analysis possible

3 Smart Physical Training Ranges

Physical training spaces of tomorrow could benefit from many types of emerging technologies. In this paper we select and review two types of technical solutions that demonstrated a capability to enrich multiple segments and aspects of training events: sensor systems or, more precisely automated sensor management systems, and virtual reality technologies.

3.1 Automated Sensor Management System

Sensor systems are already the reality of physical training spaces - the use of complex camera systems (optical sensors), for example, is not new. Those systems typically employ a number of skilled operators who manipulate camera parameters, change a viewing angle of each camera, zoom-in on action, decide what to record as information critical for future AAR and what to keep as input for a unit take-away package. With the number of sensors increasing dramatically (e.g. USMC Kilo2 training range

in Camp Pendleton has over 500 cameras), the task is becoming impossible even for a team of skilled operators, not to mention the cost of employing that team. The reality of contemporary sensor systems is that the human operators are no longer capable of managing a large data throughput and, at the same time, providing optimal performance. In order to address that situation modern camera systems need to be able to do a self-calibration, select viewing angles and zoom factors for all PTZ cameras using an optimization technique suited to the type of environment and type of performance likely to be seen in their field of view (FOV), and make smart decisions on what to record in order to reduce the data storage requirement and processing of 'empty' data. Examples of such novel systems employ a sequential Kalman Filter [3], as well as a stochastic performance metric and a constrained optimization method [4]; both methods have the same goal – to optimize the performance of optical sensors without the assistance of a human operator.

3.2 Sensor-Enabled Automated Behavior Analysis and Performance Evaluation

In the pursuit of better physical training ranges, one would like to keep everything that makes them irreplaceable and add the good features of computer-supported solutions. One way of doing this is to transform the elements of the analog world of physical training ranges into the digital domain, and in that way be able to afford the flexibility and data manipulation that is a clear advantage of the digital domain.

This type of transformation was at the very center of the Behavior Analysis and Synthesis for Intelligent Training (BASE-IT) effort [5] - the ultimate goal of the project was to greatly increase the amount of observed behavior and improve the quality of the After Action Review (AAR). An in situ network of automatically controlled PTZ cameras and personal position and orientation sensing devices was used to create dynamic three-dimensional (3D) participant models and combine them with a static 3D model of the environment [6], [7]. Sensor technology was employed to do the *sensing* part of the job – to track the movements and behavior of every individual [8], [9]. The subsequent analysis – *sense making* - was capable of deriving an automated understanding of unit actions [10], [11] ("Unit has been patrolling down Juliet Road for 15 minutes in a double column formation"), as well as identifying the incidents of team and individual performance [12] ("Fire team 1 identified without cover", "Trainee #N passed too close to doors and windows of the building #8"). Figures 2 and 3 illustrate a basic principle of both steps – *sensing* and *sense making*. After the data have been obtained, the system allowed not only viewing of the dynamic multi-dimensional participant pose tracks from any perspective in play-back mode, but also the analysis and searching of the full data set for significant events.

Once the training event is over, the take-away package for the unit is no longer a stack of DVDs that most units do not have time to watch and review, but rather a searchable database that has a full record of unit movements throughout the training range and training event, and an account of individual and team performance that can be quickly searched and interactively viewed i.e. 3D representation of the training event can be 'navigated through'. This effectively replaces multiple hours of watching

video footage with a series of quick, straight-to-the-point searches on incidents and events the unit wants to know about and examine from several different angles. Having this type of take-away package on a lightweight platform that the unit members can take with them and review in the barracks at their leisure, makes it even more flexible and certainly very attractive to the unit. This way, the value of data captured on the training range during the training event gets extended to the time after the training event. Being able to know 'who did what' and 'how many times' during any given training event, provides a great basis for directing skill remediation to the individuals and groups who need such additional training the most (example search: "Find all places where Squad #2 was bunching up").

Being able to query the performance of a unit in a single training event has great value for that unit. An additional value that has the potential to be applicable to the entire DoD service could be derived if this type of data is collected for a number of units and for an extended period of time. The data allows historical trends analysis to be performed and potentially for decisions to be made on what type of training regimen needs to be introduced to address the performance deficiencies identified in this analysis. The BASE-IT project incorporated this type of analysis as well.

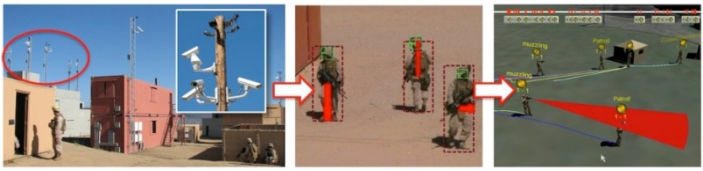


Fig. 2. System 'sensing' (*left* and *middle*) and 'sense making' (*right*)

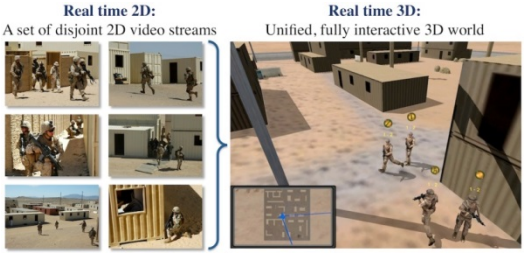


Fig. 3. Transition from a set of two-dimensional information to a unified three-dimensional world

3.3 Sensor Enabled Production Tools for the Instructors

The audience that is usually the most direct beneficiary of technical augmentations applied in training solutions is the trainees. It is less often that the instructors (trainers) are the subjects of these efforts. While automated tutors are slowly getting introduced in contemporary training solutions, human instructors are still considered irreplaceable in many aspects of skill acquisition (example: tactical decision making

where human understanding of the context plays a significant role in evaluation and subsequent change of instruction during the training event). Providing the instructors with the tools that allow them to be more productive has the potential to increase the overall effectiveness of the training event. The type of information they would be presented with during a training event to use as a basis for closed-loop feedback decisions (real-time change and re-direction of training event), and after the event as a basis for AAR, is the very same information derived from the sensor system, automated behavior analysis and automated performance evaluation [6], [7].

3.4 Augmentation of Real Spaces with Synthetic (Virtual) Elements

Physical training ranges get very close to the look-and-feel of operational environments, however they are different from operational environments in several aspects. One difference is that some cues provided in training environments are still notional – the artifacts are not actually present (visible) to the trainees, and are only imaginary. The trainees are expected to integrate that information in their situational awareness and act as if those artifacts are actually present. Examples of such cues are representations of air assets (e.g. planes) and vehicles approaching their locations, explosions, and humans – artifacts that would complete a description of the particular situation the trainees are expected to act upon.

In order to increase the level of realism, and ensure the trainees act as if they believe those artifacts are present in their immediate physical environment, virtual reality (VR) and augmented reality (AR) solutions have been suggested and tested in different training setups. The Dismounted Augmented Reality Training System (DARTS) research project demonstrated the ability to inject virtual entities (tanks, targets, opposing forces) inside real environments in real time using a man-worn, non-tethered augmented reality system [13]. The Future Immersive Training Environment JCTD research project demonstrated a training solution with an immersive training environment for small-unit ground forces [14], [15]. Similar to DARTS, the trainees used individual-worn augmented reality systems with see-through, head-mounted display. Another form of augmentation used in the same effort were VR projections on the walls representing 'through the window' view (Vista Screen) of a distant virtual terrain, roads, and neighboring village, as well as the virtual humans i.e. role players in different rooms acting as a part of a particular training scenario.

The test runs with actual units produced feedback about the way trainees' treated 2D projections of virtual humans: the trainees observed and acted on animated scenes, however as soon as a real human would enter the room, the trainees' attention would shift from the 2D virtual humans to the real human. In other words, the physical world was given higher priority and 'images' on the wall got less attention. This observation was the inspiration for a line of research initiated by our team [16]. The research effort included work with tangible virtual humans – physical 3D models of humans with animated images of human faces being projected on them; the technique is known as the shader-lamp approach [17]. Our hypothesis was that 3D virtual human will be more effective representations than 2D virtual humans and our results are very promising (a paper with a full account of the study results is in preparation). This is not to

say that 2D virtual humans do not have their own role - in some situations they will continue to be used as an effective part of the human 'landscape'. A good example of such situations are training sessions of virtual rowing where the presence of virtual competitors in the visual field of view of the trainees was found to influence the performance of the rowers [18]. Figure 4 provides an illustration of several types of role players: (a) real people in the USCM Immersion Infantry Trainer (IIT), (b) 2D projection on the wall in IIT [19], (c) animatronic characters in IIT (Garner Holt Productions Inc. and Lockheed Martin), and (d) 3D tangible humans [16]. We believe that future physical training ranges will use what we call a human tapestry – an eclectic mix of different forms and types of virtual humans, each type fulfilling the role they are best suited for. The ongoing and future research efforts will need to determine what those roles are and what is the best way to incorporate them in training scenarios. It is important here to note that the augmentation of physical training ranges does not have to happen only in the visual sensory domain – training ranges can have augmentation with auditory (sound) as well as olfactory sensory data (odors). Both of those have been successfully tested and deployed in the IIT at Camp Pendleton.

(a) (b) (c) (d)

Fig. 4. Role players: (a) live, (b) two-dimensional projection of role players, (c) animatronic character, (d) tangible three-dimensional virtual human (shader-lamp projection)

4 Future Generations of Physical Training Environments

Training on physical ranges is and will continue to play an important role in training of the military – it is hard to expect any change in that regard for many elements of individual and team skills. However, what is very likely to change is the fidelity and variety of information presented to the training force, as well as the way training data gets collected, manipulated and made available to the trainees for study and analysis.

We envisage that in that process the emphasis will inevitably be placed on a greater level of automation in data collection and analysis, with the objective being to serve the needs of all users (trainees, instructors, operators). The form of training solutions likely to be present in the future are an organic, eclectic mix of technologies, systems, tools, sensor solutions and training approaches that most effectively support the training objectives in a given training environment and for a particular training audience. As has been the case so far, one specific mix will work very well for the needs of small units operating in building and room clearing, and a different set will be needed for larger cordon and search operations.

We foresee that in the future a variety of sensors will be integrated in a unified sensor system solution, where each sensor type could also be used to strengthen and

optimize the performance of another sensor. An example of such a synergetic mix would be the use of sensors detecting movement, with that information being used as a predictor for 'what is coming and how fast' to the FOV of nearby PTZ cameras and microphones. One could also imagine that, in addition to an automated sensor management approach, a mix of an automated approach with occasional human intervention may become one of the modes of operation (humans making judgments in situations that are advantageous to human situational awareness and human reasoning). Different types of sensors and instrumentation systems are already in use, like laser tactical engagement simulators OneTESS, ITAS-TESS and MILES XXI [20] and a variety of stationary and mobile target systems. A synergetic combination of existing sensor solutions strengthened with new types of sensor data would amplify the value of investments already made in acquired range instrumentation systems.

Similar to the capture of trainees' movements in physical space, the ability to acquire and quickly analyze audio signals (shouts, communication, grunts) and provide that as contextual information tied to each individual inside a 3D dynamic model, and to do automated behavior analysis and performance evaluation with that multi-layered information would be very valuable. The same physical movements and actions may be understood very differently if shouts and commands given at critical points during those actions were heard and understood by the instructors (evaluators). Therefore it will be important to obtain as comprehensive understanding about the event as possible, and analyze it for the benefit of all users. The experiences acquired in military training systems are applicable to several other domains, like training of fire fighters, police forces, port and airport security, as well as a range of sport training situations – in short all types of human activity where it is important to track and provide information about the progression of participants' skill mastery.

Acknowledgments. The author would like to acknowledge the following agencies and individuals for their continuing support: the sponsor Office of Naval Research (ONR), the Program Manager for Training Systems (PM TRASYS) and Training and Education Command (TECOM), leadership and instructors of Tactical Training Exercise Control Group (TTECG), Twentynine Palms and Kilo Two, Camp Pendleton, as well as many USMC units and individuals who supported different elements of our research efforts.

References

1. The Strategic Plan for the Next Generation of Training for the Department of Defense, Office of the Under Secretary of Defense (Personnel & Readiness) (2010)
2. Sadagic, A., Darken, R.: Combined Arms Training: Methods and Measures for a Changing World. In: NATO Workshop Virtual Media for Military Applications, US Military Academy, West Point (2006)
3. Deutsch, B., Niemann, H., Denzler, J.: Multi-step Active Object Tracking With Entropy Based Optimal Actions Using the Sequential Kalman Filter. In: IEEE International Conference on Image Processing, vol. 3 (2005)

4. Ilie, A., Greg Welch, G.: On-Line Control of Active Camera Networks for Computer Vision Tasks. In: Proceedings of 5th ACM/IEEE International Conference on Distributed Smart Cameras, Ghent, Belgium (2011)
5. Behavior Analysis and Synthesis for Intelligent Training (BASE-IT), MOVES Institute, Naval Postgraduate School, http://www.movesinstitute.org/base-it
6. Sadagic, A., Welch, G., Basu, C., Darken, C., Kumar, R., Fuchs, R., Cheng, H., Frahm, J.M., Kölsch, M., Rowe, N., Towles, H., Wachs, J., Lastra, A.: New Generation of Instrumented Ranges: Enabling Automated Performance Analysis. In: Proceedings of Interservice/Industry Training, Simulation, and Education (I/ITSEC) Conference (2009)
7. Sadagic, A., Kölsch, M., Welch, G., Basu, C., Darken, C., Wachs, J.P., Fuchs, H., Towles, H., Rowe, N., Frahm, J.-M., Guan, L., Kumar, R., Cheng, H.: Smart Instrumented Training Ranges: Bringing Automated System Solutions to Support Critical Domain Needs. In: The Journal for Defense Modeling and Simulation, JDMS (accepted for printing in 2013)
8. Viola, P., Jones, M.: Robust Real-time Object Detection. In: 2nd Int. Workshop on Statistical and Comp. Theories of Vision, Vancouver (2001)
9. Wachs, J., Goshorn, D., Kölsch, M.: Recognizing Human Postures and Poses in Monocular Still Images. In: International Conference on Image Processing, Computer Vision, and Pattern Recognition - IPCV (2009)
10. Cheng, H., Yang, C., Han, F., Sawhney, H.: HO2: A new feature for multi-agent event detection and recognition. In: Computer Vision Pattern Recognition Workshop (2008)
11. Cheng, H., Kumar, R., Basu, C., Han, F., Khan, S., Sawhney, H., Broaddus, B., Meng, C., Sufi, A., Germano, G., Kölsch, M., Wachs, J.: An Instrumentation and Computational Framework of Automated Behavior Analysis and Performance Evaluation for Infantry Training. In: Proceedings of Interservice/Industry Training, Simulation, and Education, I/ITSEC (2009)
12. Rowe, N., Houde, J., Kölsch, M., Darken, C., Heine, E., Sadagic, A., Basu, C., Han, F.: Automated Assessment of Physical-Motion Tasks for Military Integrative Training. In: 2nd International Conference on Computer Supported Education, Valencia, Spain (2010)
13. Dean, F., Jaszlics, S., Stilson, R., Sanders, S.: Augmented Reality: Enabling Component for Effective Live Virtual Constructive Integration. In: Virtual Media for Military Applications (2006)
14. Future Immersive Training Environment Joint Capability Technology Demonstration, Operational Demonstration 2. Independent Assessment Report (2011)
15. Ross, W.A., Kobus, D.A.: Case-Based Next Generation Cognitive Training Solutions. In: Proceedings of Interservice/Industry Training, Simulation, and Education (I/ITSEC) Conference (2011)
16. 3D Display and Capture of Humans for Live-Virtual Training, MOVES Institute, Naval Postgraduate School, http://www.movesinstitute.org/MovesVirtualHumans
17. Raskar, R., Welch, G., Low, K.-L., Deepak Bandyopadhyay, D.: Shader Lamps: Animating Real Objects With Image-Based Illumination. In: Eurographics Workshop on Rendering (2001)
18. Wellner, M., Sigrist, R., Riener, R.: Virtual Competitors Influence Rowers. Presence 19(4), 313–330 (2010)
19. Combat Hunter Action and Observation Simulation (CHAOS), Institute for Creative Technology, University of Southern California, http://ict.usc.edu/prototypes/chaos/
20. U.S. Army PEO STRI: Product Manager Live Training Systems – PM LTS, http://www.peostri.army.mil/PM-TRADE/lts.jsp

A Study on Application of RB-ARQ Considering Probability of Occurrence and Transition Probability for P300 Speller

Eri Samizo, Tomohiro Yoshikawa, and Takeshi Furuhashi

Nagoya University, Nagoya 464-8603, Japan

Abstract. Brain-Computer Interfaces (BCIs) control a computer or a machine based on the information of the signal of human's brain. P300 speller is one of the BCI communication tools, which uses P300 as the feature quantity and allows users to select letters just by thinking. Because of the low signal-to-noise ratio of the P300, signal averaging is often performed to improve the spelling accuracy instead of the degradation of the spelling speed. In texts, there is variability in occurrence probabilities and transition probabilities between lettersDThis paper proposes P300 speller considering the occurrence probabilities and the transition probabilities as the prior probabilities in RB-ARQ. It shows that the spelling speed and then the Utility were improved by the proposed method comparing with the conventional method.

1 Introduction

Brain-Computer Interface is the system that controls a computer or a machine based on the information of signals from human's brain[1]. It is expected to be developed as a communication tool for seriously paralyzed patients like those with amyotrophic lateral sclerosis (ALS). Electroencephalogram (EEG) is most likely used for BCIs because it is noninvasive and inexpensive. P300 speller that is first introduced by Farwell et al. is one of the communication tools using P300 as a feature[2]. P300 is one of the event-related potential (ERP) and it is elicited when a stimulus that a user attends to is provided. A user can choose and input letters just by his/her thoughts using P300 speller. It generally uses a letter matrix interface with visual stimulus. Each row or column flashes in random order one by one for a certain times. While they are flashing, the user concentrates on the desired letter by counting how many times it flashes. Thereby, P300 is elicited when the row or column that contains the desired letter is flashed. Then the system discriminates the letter that includes P300 most likely as the target one.

However, signal-to-noise ratio of the P300 is small. Thus, averaging signals is needed[3][4], which improves the spelling accuracy instead of degrading the spelling speed. Practically, it is needed to input letters correctly in a short time to reduce user's burden. Conventionally, the number of flashing times, i.e., the number of stimuli, is fixed. To reduce the number of stimuli, Reliability-Based

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 727–733, 2013.

Automatic Repeat reQuest (RB-ARQ) has been proposed[5]. It is shown that RB-ARQ can reduce spelling speed with keeping spelling accuracy[6].

In RB-ARQ, the prior probability, the likelihood of each letter to be the target before the presentation of stimuli, is set equally for all letters. On the other hand, there is variability in occurrence probabilities and transition probabilities between letters in texts. In the area of understanding texts or voice recognition, the transition probabilities between letters are used for letter correction or the support of input and recognition[7][8].

In this paper, we propose a new P300 speller that considers the ocurrence probabilities and the transition probabilities between letters as the prior probability in RB-ARQ. The experiments are done by three subjects with Japanese interface of P300 speller and the result shows the improvement of spelling speed and then the Utility, which is the performance index of spelling considering accuracy and discrimination time at once, by the proposed method comparing with the conventional one.

2 Reliability-Based Automatic Repeat reQuest

RB-ARQ is a method that presents stimuli randomly and sets the number of stimuli dynamically based on the maximum posterior probability[5][6]. Suppose $\boldsymbol{x}_t$ denotes a feature vector from EEG data at time t, and let $X_T = \{\boldsymbol{x}_t | t = 1, 2, ..., T\}$ be a set of data at time T, the posterior probability at time T can be calculated as follows:

$$P(k|X_T) = \frac{P(k)\prod_t p(\boldsymbol{x}_t|k)}{\sum_{l\in K} P(l)\prod_t p(\boldsymbol{x}_t|l)} \tag{1}$$

In this equation, let K be a set of candidate letters and $k \in K$. And $P(k)$ is the prior probability that $\boldsymbol{x}$ belongs to label k before the stimulus presentation, and they are set equally. The posterior probability is obtained by multiplying the prior probability and likelihood. Maximum posterior probability at time T is defined as Eq.(2) using the posterior probability $P(k|X_T)$.

$$\lambda_T = \max_k P(k|X_T) \tag{2}$$

The maximum posterior probability is equivalent to the discrimination accuracy, which can be regarded as the reliability of data. λ is set as the threshold of reliability, and a user keeps thinking until λ_T becomes larger than λ.

3 Proposed Method

As mentioned above, the prior probability of RB-ARQ is set equally to every letter in the conventional method. This paper proposes a method to consider the occurrence probability and the transition probability of letters in text as the prior probability. Transition probability is the frequency of a letter in texts after

the given preceding letter(s), and it is given by the occurrence rate of N-gram character in an enormous quantity of text data. N-gram is every contiguous sequence of n characters in a given text[9]. Therefore, the prior probability is defined as below with n=1,2,...

$$P(X_i) = \frac{N(X_i)}{\sum_{l \in K} N(X_i)} \qquad (n = 1) \tag{3}$$

$$P(X_i | X_{i-n+1}^{i-1}) = \frac{N(X_{i-n+1}^{i})}{N(X_{i-n+1}^{i-1})} \qquad (n \geq 2) \tag{4}$$

Let X_i^j be a part of string from ith letter to jth letter in the character string $X_1 X_2, ... X_M$. $P(X_i | X_{i-n+1}^{i-1})$ is the conditional probability that ith letter becomes X_i when a string from $\{i-(n-1)\}$th letter to $(i-1)$th letter is given. $N(X_i^j)$ denotes the occurrence frequency of a string from ith letter to jth letter. When n is 1, the prior probability is simply represents the probability of the occurrence of each letter, and this paper calls iteUni-gram.fWhen n is 2 or 3, they represent the transition probability between letters. This paper calls themeBi-gramfandeTri-gram,frespectively. Using these probabilities, it is expected to improve the performance of inputting text in RB-ARQ. It is thought that the time until the posterior probability exceeds the threshold λ becomes shorter, because the letters with high occurrence rate have high prior probability.

4 Experiment

4.1 Data Description and Preprocessing

This experiment used a recorded dataset which contained EEG data measured by three subjects (Sub A, Sub B and Sub C) performed the P300 speller. EEG data was recorded with sampling frequency of 1000Hz using Polymate AP216 (Digitex lab. co., ltd., Tokyo), from 5electrodes: Fz, Cz ,Pz, O1 and O2, referenced to the linked ears, A1 and A2 (Fig.1). The stimulus onset asynchrony (SOA) was 175 ms: each stimulus was presented for 100 ms with an inter-stimulus interval (ISI) of 75 ms. In this experiment, the 7-by-10 letter matrix interface containing Japanese characters shown in Fig.2 was employed. An input of one letter consisted of ten sequences, while one sequence contained 17 (10 rows and 7 columns) stimuli. Then, the EEG signals were down-sampled to 20Hz, 14 data points corresponding to 0s to 0.65s after each stimulus was extracted. The extracted data were classified using Linear Discriminant Analysis (LDA). 20 letters were utilized for the learning session.

In this interface, <> is used when the user wants to input small letters. For example, when the user wants to input <>, he/she needs to select <> before <>. And <BS> means backspace which deletes the preceding letter. The prior probability employed in the proposed method was calculated based on the web corpus of Japanese[10]. In the calculation of Eq.(3)(4), sonant marks and p-sounds were regarded as one character, and the prior probability of <> was calculated using the number of the appearance of small letters.

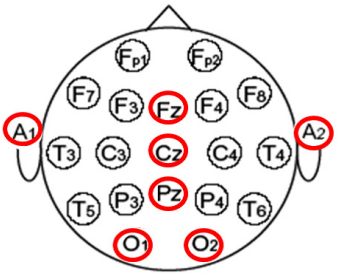


Fig. 1. Used electrodes

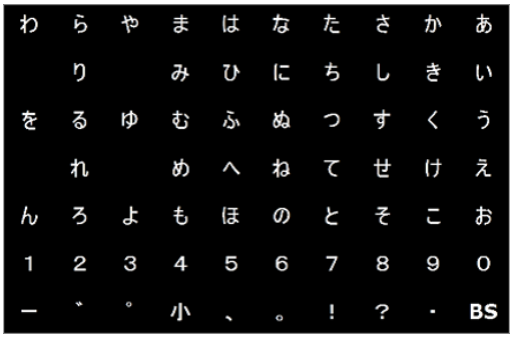


Fig. 2. User-interface

4.2 Experimental Settings and Performance Index

The experiment employed three long sentences, which had about 200 letters extracted from Web blog, essay and novel, respectively. They were inputted for 200 times in each test, and the results were averaged.

When non-target letter was inputted, that is, discrimination result was wrong, <BS> would be selected at the next target to input full sentence correctly. The threshold of RB-ARQ was set to 0.9, 0.95 and 0.99. We conducted the following two experiments.

Exp.1: We set the prior probability by Uni-gram, Bi-gram and Tri-gram and compared with the conventional method, the prior probability was set equally caled eEqual.'

Exp.2: We set the prior probability by Uni-gram, Bi-gram and Tri-gram, and when <BS> and a letter except for <BS> were repeated, we set every prior probability equal.

This paper employed the performance index for the comparison as the average of the discrimination accuracy, the number of stimuli and the discrimination (input) time per a letter. Each performance index is determined below.

$$Discrimination\ accuracy = \frac{\sharp\ of\ correct\ letters}{\sharp\ of\ inputted\ letters} \tag{5}$$

$$\sharp\ of\ stimuli\ per\ a\ letter = \frac{\sharp\ of\ all\ stimuli}{\sharp\ of\ inputted\ letters} \tag{6}$$

$$Discrimination\ time\ per\ a\ letter = \sharp\ of\ stimuli\ per\ letter \times SOA \tag{7}$$

This paper also uses gUtility[11]hdefined in Eq.(8) to evaluate the accuracy and the discrimination time at once.

$$U = \frac{(2P-1)\log_2(C-1)}{d} \tag{8}$$

where C is the number of classes (in this experiment, N=70, 10 rows 7 columns), P is the accuracy, and d is the discrimination time per a letter. Note that if P <0.5, U=0. Utility corresponds to the information transfer rate when the spelling is done perfectly by using <BS> that can delete incorrect characters. Thus, it is thought to be a practical performance measure for the P300 speller.

5 Result

Table 1 shows the discrimination accuracy, the number of stimuli and the discrimination time in Exp.1, and Table 2 shows those in Exp.2. Figure 3 shows the Utility in Exp.1, and Fig.4 shows that in Exp.2. In these figures, the values on the horizontal axis mean the thresholds of RB-ARQ and vertical axis shows the value of Utility.

Table 1. Performance indexes in Exp.1

	Threshold	Equal	Uni-gram	Bi-gram	Tri-gram
Accuracy	0.9	0.81	0.794	0.818	0.65
	0.95	0.869	0.87	0.891	0.819
	0.99	0.918	0.933	0.95	0.939
♯ of Stimuli	0.9	78.9	68	55.2	41.9
	0.95	90.6	80	66.2	57.7
	0.99	110.8	100.6	85.4	83
Time[s]	0.9	13.8	11.9	9.7	7.3
	0.95	15.9	14	11.6	10.1
	0.99	19.4	17.6	14.9	14.5

Table 1 shows that the accuracy of Uni-gram and Bi-gram were almost equal or better than the conventional method (Equal), while, the number of stimuli of Uni-gram and Bi-gram were smaller than Equal. On the other hand, the number of stimuli of Tri-gram at threshold 0.9 was also reduced, however, the accuracy decreased at the same time. Especially in Tri-gram, the prior probability widely varied depending on the next selectable letters comparing with other methods. Threfore, the discrimination time was largely decreased when the letter with high prior probability was selected as the target letter. On the other hand, when the letter with low prior probability was chosen as the target and the threshold in RB-ARQ was low, a non-target letter with high prior probability was tend to be selected incorrectly. When a non-target letter was inputted, the subject needed to input <BS> for the correction. Then, the discrimination accuracy became low because of the repetition of inputting <BS> and a non-target letter. Thus in the threshold higher than 0.9, the accuracy of Tri-gram improved because this repetition happened less frequently. As the result, Fig.3 shows that the

Table 2. Performance indexes in Exp.2

	Threshold	Equal	Uni-gram	Bi-gram	Tri-gram
Accuracy	0.9	0.81	0.797	0.805	0.777
	0.95	0.869	0.869	0.88	0.859
	0.99	0.918	0.933	0.944	0.941
♯ of Stimuli	0.9	78.9	64.7	58.8	54.1
	0.95	90.6	76.5	69.8	63.6
	0.99	110.8	98.9	91	83.2
Time[s]	0.9	13.8	11.3	10.3	9.5
	0.95	15.9	13.4	12.2	11.1
	0.99	19.4	17.3	15.9	14.6

performance in Utility of Uni-gram was better than Equal, Bi-gram was superior to Uni-gram and Tri-gram at threshold 0.95 or 0.99 was better than Bi-gram, while that of Tri-gram at threshold 0.9 was the worst.

On the other hand, Table 2 shows the improvement of the accuracy of Tri-gram at every threshold in Exp.2, which has the avoidance of the repitition by setting every prior probability equal. In this experiment, test sentences consisted of about 200 letters including a lot of particles which had poor connection with next letter. Thus, there were many times that the letter with low prior probability was to be selected as the target. In Exp.1, the prior probability was set to N-gram at all times, therefore, the repetition in Tri-gram was affected largely, which was improved by using the equal prior probability effectively with N-gram. In the comparison of Equal with Tri-gram, the discrimination time was shorten from 4 to 5 seconds per a letter. Thus in Fig.4, the performance in Utility of Tri-gram was superior to Bi-gram at every threshold.

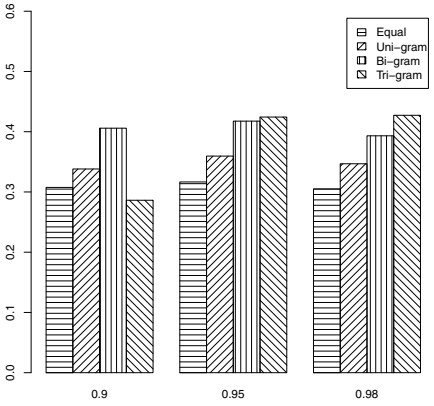


Fig. 3. Utility in Exp.1

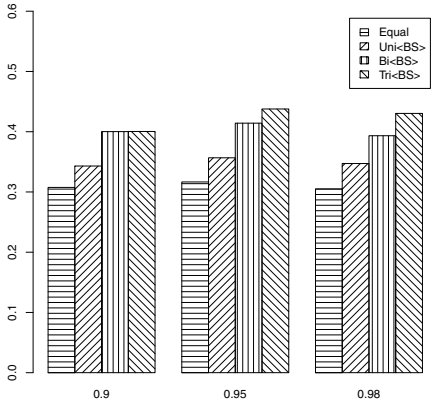


Fig. 4. Utility in Exp.2

There were the significant differences in Utility between the conventional method and every proposed method by the paired t-test at the significant level of $\alpha = 0.017$ (0.05/3; Boneferroni correction) considering multiple comparison. This result showed that considering the occurrence probability and the transition probability by the proposed method improved the inputting performance.

6 Conclusion

This paper proposed P300 speller that considering the occurrence probabilities and the transition probabilities between letters as the prior probability in RB-ARQ. The experiments were done by three subjects with Japanese interface of P300 speller and the result showed the improvement of spelling speed keeping high accuracy by the proposed method comparing with the conventional one. We will more investigate the proposed method through the online experiment.

References

[1] Wolpaw, J., Birbaumer, N., McFarland, D., Pfurtscheller, G., Vaughan, T., et al.: Brain-computer interfaces for communication and control. Clinical Neurophysiology 113(6), 767–791 (2002)
[2] Farwell, L.A., Donchin, E.: Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and Clinical Neurophysiology 70(6), 510–523 (1988)
[3] Hoffmann, U., Vesin, J.M., Ebrahimi, T., Diserens, K.: An efficient p300-based brain–computer interface for disabled subjects. Journal of Neuroscience Methods 167(1), 115–125 (2008)
[4] Scherer, R., Muller, G.R., Neuper, C., Graimann, B., Pfurtscheller, G.: An asynchronously controlled eeg-based virtual keyboard: improvement of the spelling rate. IEEE Transactions on Biomedical Engineering 51(6), 979–984 (2004)
[5] Takahashi, H., Yoshikawa, T., Furuhashi, T.: A study on application of reliability based automatic repeat request to brain computer interfaces. In: Proc. 15th Int. Conf. Neural Information Processing, pp. 1013–1020 (2009)
[6] Kaneda, Y., Takahashi, H., Yoshikawa, T., Furuhashi, T.: A study on application of reliability-based automatic repeat reQuest to p300 speller. IEICE Technical Report 109(280), 19–22 (2009)
[7] Kukich, K.: Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR) 24(4), 377–439 (1992)
[8] Ullmann, J.R.: A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. The Computer Journal 20(2), 141–147 (1977)
[9] Cavnar, W., Trenkle, J., et al.: N-gram-based text categorization, pp. 161–175, Citeseer, 48113 (1994)
[10] `http://s-yata.jp/corpus/nwc2010/ngrams`
[11] Dal Seno, B., Matteucci, M., Mainardi, L.T.: The utility metric: a novel method to assess the overall performance of discrete brain–computer interfaces. IEEE Transactions on Neural Systems and Rehabilitation Engineering 18(1), 20–28 (2010)

Improvement of Sensory Stabilization and Repeatability of Vibration Interface for Distance Presentation

Yuki Sampei[1], Takayuki Tanaka[1], Yuki Mori[2], and Shun'ichi Kaneko[1]

[1] Hokkaido University, Kita 14 Nishi 9,Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
sampei@ssc.ssi.ist.hokudai.ac.jp
[2] Riken, Nagoya, Japan

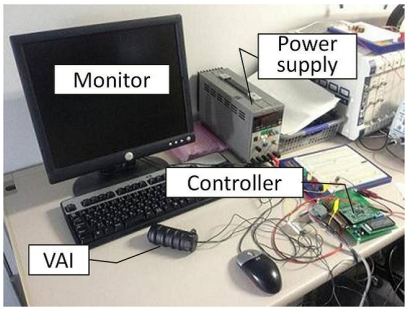
Abstract. We have developed a vibration alert interface (VAI) that provides information through various vibration patterns. In our previous studies, we designed the VAI and its vibration patterns to provide analog-like information to users such as distance to obstacles. Precise information recognition requires correct perception of vibration patterns. However, various disturbances can affect perception of vibrations, causing users to perceive similar vibrations as being different. We therefore proposed the *relative vibration sense presentation* method to avoid disruption of the vibration sense. In this paper, we experimentally show that this method improves the repeatability of vibration sensation. We also propose a vibration presentation model for drivers to correct perception gaps due to the application and surroundings of VAI. We evaluate the proposed model through experimentation.

1 Introduction

We developed the vibration alert interface (VAI) to provide information through varying vibration patterns. Vibration is an effective method for conveying information to individual users. In previous studies, we designed VAI and its vibration patterns to provide analog information such as distance to obstacles [1]. Conventional vibration devices convey information by turning vibrations on or off [2][3]. We found that a higher frequency vibration motor can convey greater vibration strength to users.

Precise information recognition requires correct perception of vibration patterns. However, various disturbances can affect perception of vibrations, causing users to perceive similar vibrations as being different [4][5]. One of the things which give information by vibration is a sound. A sound is defined as follows [6]. A phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system. The phonemes cannot be defined acoustically and they are a set of abstractions. It state that it isn't sound but a sound difference, i.e., contrast, which should be perceived in a sound system [7]. We therefore developed a new presentation method for vibration that avoids disruption of the vibration sense. The proposed *relative vibration*

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 734–743, 2013.



(a) Experimental setup and VAI.

(b) State of participant and VAI.

Fig. 1. The vibration alert interface (VAI)

sense presentation (RVSP) method alternately presents a presentation vibration frequency f_p and a constant base vibration frequency f_b.

In this paper, we experimentally show that RVSP improved repeatability of vibration sensations. Further, we propose a vibration presentation model for drivers to correct perception gaps due to the application and surroundings of VAI. We evaluate the proposed model through experimentation.

2 Vibration Alert Interface

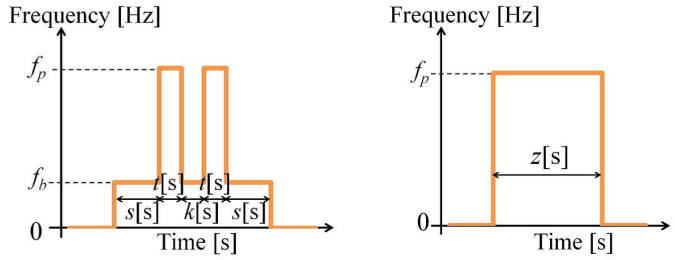
2.1 Experimental Systems and Conditions

Fig. 1(a) shows the structure of the VAI experimental device. The vibrator is a cylindrical plastic object containing a small vibrating motor. Participants held it, and vibration frequency was controlled through an H8 microcomputer and PC-based motor driver. The participants in the experiment were four healthy adult men and women age 21 to 22 (21.8 average). To ensure only tactual judgment, participants wore eye masks and listened to white noise via headphones (Fig. 1(b)).

According to Weber–Fechner's law, sensations are perceived in proportion to the logarithm of the stimulation [8][9]. The amplitude of VAI vibration used in this research is independent of the frequency, so we can define the quantity of user vibration perception as

$$E = k_f \log f + C, \tag{1}$$

where f is the vibration frequency, k_f is the gain, and C is a constant. Since the energy of vibration is proportional to the logarithm of the frequency, the magnitude of the energy is controllable by controlling the frequency.



(a) The relative vibration sense presentation (RVSP) method.

(b) The absolute vibration sense presentation (AVSP) method.

Fig. 2. Definitions of the presentation methods ($f_b > 0$)

2.2 Relative Vibration Sense Presentation Method

Factors related to changes in vibration perception for similar vibrational frequencies include the contact area and changes in the grip force. In this research we focus on the existence of C, the unknown constant in Eq. (1). In conventional presentation methods, since the presentation vibration f_p [Hz] is given after a state of no vibration (0 [Hz]), C varies with every presentation. We predict that the dampening of vibration sense is due to the instability of C.

We therefore propose the *relative vibration sense presentation* (RVSP) method to improve the repeatability of the vibration sense provided by the VAI. RVSP alternately presents a presentation vibration frequency f_p and a constant base vibration frequency f_b. It aims at reducing the influence of C in Eq. (1) by presenting ΔE, which is the difference between the quantity of sense by presentation vibration E_p and the quantity of sensation by base vibration E_b. k_f and C are assumed to be fixed by continuous oscillating presentation. This with Eq. (1) gives Eq. (2). We consider that users can perceive correct information from f_p in comparison with f_b by eliminating C, as in Eq. (2).

$$\Delta E = E_p - E_b = k_f(\log f_p - \log f_b) \tag{2}$$

We perform an experimental investigation to determine if RVSP improves reproducibility of vibration sense more than does the *absolute vibration sense presentation* (AVSP) method, which does not use a base vibration frequency. Fig. 2(a) shows the RVSP vibration pattern, and Fig. 2(b) shows the pattern for AVSP. In that figure, t is the presentation time of the presentation vibration in RVSP, s and k are the presentation times of the base vibration in RVSP, and z is the presentation time of the presentation vibration in AVSP. A cycle of a given vibration in RVSP is called a presentation vibration cycle.

Although $\log f$ would be infinitely large in Eq. (1) at times where $f = 0$ [Hz], there are dead zones in human perception of stimuli, and the threshold of the vibration frequency changes with the equipment used. We consider that there is no

quantity of vibration sense in states where VAI vibrates at vibration frequencies less than f_0. Eq. (1) is therefore redefined as follows:

$$E = \begin{cases} 0 & \text{if} \quad f < f_0 \\ k_f \log f + C & \text{otherwise} \end{cases} \tag{3}$$

3 Verification of the Validity of RVSP

3.1 Comparison of RVSP and AVSP

We conducted experiments that compare RVSP with AVSP to verify the validity of RVSP. Since RVSP needs the existence of the base vibration, which has the vibration frequency more than fixed, we sets it to $f_b = 100$ [Hz] and advances verification.

The parameters for RVSP were set as $s = 2$, $t = 1$, and $k = 1$ (Fig. 2). This vibration pattern is called *rel* below. The parameter of AVSP were set as $z = 5$. This vibration pattern is called *abs* below. The presentation vibration frequencies were set to 100, 115, 130, 145, and 160 [Hz]. Participants reported perceived vibration strength v_s as an integral value.

Experimental procedures were as follows:

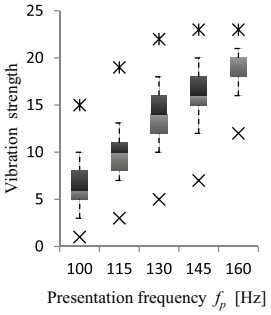
(1) Participants grasped the VAI vibrating at $f_p = 100$ [Hz], and were told to classify perception of this vibration as strength 5.
(2) Next, they were presented with a new vibration strength, $f_p = 160$ [Hz], which they were told to classify as vibration strength 20.
(3) Finally, participants were presented with vibration at a random frequency, and reported the perceived vibration strength.
(4) Steps (1)–(3) were repeated five times.

We repeated this process ten times per day for each participant, with short breaks between presentations. This was repeated over six days, resulting in 60 data points regarding vibration strength perception for various frequencies for each participant. The experiment that presents *rel* was carried out after completion of the experiment that presents *abs*.

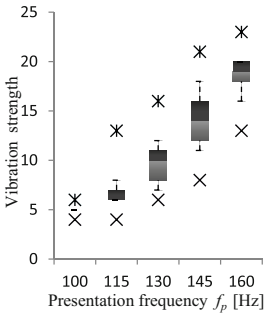
Fig. 3 shows the results for all participants. As the box plots indicate, the standard deviation for *rel* was smaller than for *abs* at all frequencies (Table 1). Comparing results between participants, the same result was obtained in 17 out of 20 pairs of data groups (four participants, five frequencies each). This confirms that the variation in vibration strength was small under RVSP, and thus that RVSP with $f_b = 0$ is effective.

3.2 Presentation Vibration Cycle

We changed the presentation vibration cycle of RVSP, and carried out other experiments. The parameters of RVSP were set as $s = 1$, $t = 1$, $k = 1$, and



(a) *abs*.



(b) *rel*.

Fig. 3. Comparing *rel* and *abs* in terms of repeatability

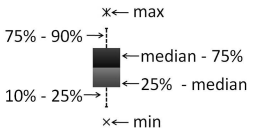


Fig. 4. How to read our box plot

Table 1. Comparing *rel* and *abs* in terms of standard deviation and improvement rate

Presentation vibration frequency [Hz]	Standard deviation		Improvement rate[%]
	abs	*rel*	*absrel*
100	2.56	0.14	94.3
115	2.99	1.12	62.4
130	3.13	1.94	37.9
145	2.82	2.55	9.70
160	2.08	1.72	17.6
Average	2.71	1.50	44.4

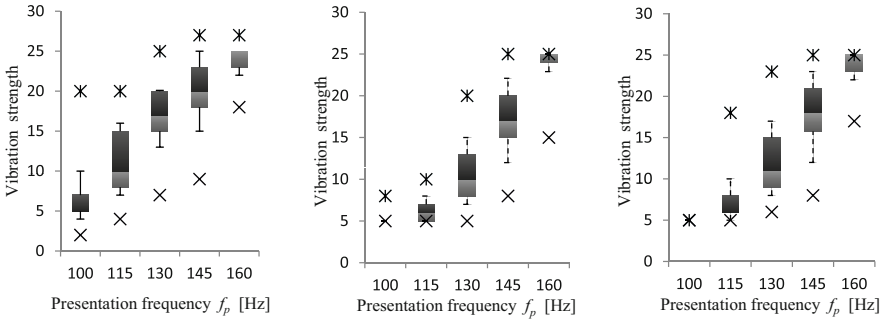
$s = 2$, $t = 2$, $k = 2$ (Fig. 2). These vibration patterns are respectively called *rel*-1 and *rel*-2 below. AVSP parameters were set as $z = 3$. This vibration pattern is called *abs*-3 below. The presentation vibration frequencies were set to the same five levels as in the previous section.

Experimental procedures were as follows:

(1) Participants grasped the VAI vibrating at $f_p = 100$ [Hz], and were told to classify perception of this vibration as strength 5.
(2) Next, they were presented with a new vibration strength, $f_p = 160$ [Hz], which they were told to classify as vibration strength 25.
(3) Finally, Participants were presented with vibration at a random frequency, and reported the perceived vibration strength.
(4) Steps (1)–(3) were repeated five times.

We repeated this process 15 times per day for each participant, with short breaks between presentations. This was repeated over 3 days, resulting in 45 data points regarding vibration strength perception for various frequencies for each participant. We carried out these experiments in the order *abs*-3, *rel*-1, then *rel*-2.

Fig. 5 shows the results for all participants. Standard deviations of *rel*-1 and *rel*-2 were smaller than that of *abs*-3 at frequencies 100 and 115 [Hz], but larger at frequencies 130, 145, and 160 [Hz] (Table 2). In between-participant comparisons, however, *rel*-1 was smaller than *abs*-3 in 17 of 20 pairs, and *rel*-2 obtained the same result in 18 of 20 pairs. We thus found that the variation in vibration strength was small under RVSP for each participant. The above analysis shows that in each participant *rel*-1 and *rel*-2 suppressed variation in the vibratory sense, as compared with *abs*-3.



(a) *abs*-3. (b) *rel*-1. (c) *rel*-2.

Fig. 5. Comparison of *rel*-1, *rel*-2, and *abs*-3 for differences in vibration perception

Table 2. Comparison of *rel*-1, *rel*-2, and *abs*-3 standard deviation and improvement rate

Presentation vibration frequency [Hz]	Standard deviation			Improvement rate[%]	
	abs	*rel*-1	*rel*-2	*abs*-3*rel*-1	*abs*-3*rel*-2
100	2.47	0.25	0.00	90.0	100
115	3.70	1.47	1.93	60.2	48.0
130	3.33	3.94	3.58	-4.91	-7.51
145	3.65	3.74	3.98	-2.51	-9.19
160	1.66	1.69	1.67	-1.80	-0.918
Average	2.95	2.13	2.23	28.2	26.1

4 Perceptual Model to Convert Vibration Stimulation to Distance Perception

4.1 Systems

We assume two hypotheses as in Fig. 6 concerning the mechanism by which humans recognize vibratory stimulation as distance perception.

(I) The relation H_1 between vibratory stimulation and vibration strength is independent of the VAI application.
(II) The relation H_2 between vibration strength and distance perception is dependent on the VAI application.

We examine the relations H_1 and H_2 during driving to develop VAI-C, which applies VAI to support drivers with information such as distance to obstacles. We then verify the above-mentioned hypotheses. The influence on these relations by driving speed is experimentally investigated with a driving simulator. Finally, we confirm that VAI-C presents stable distance perception, independent of driving speed. Fig. 7 shows the experimental setup.

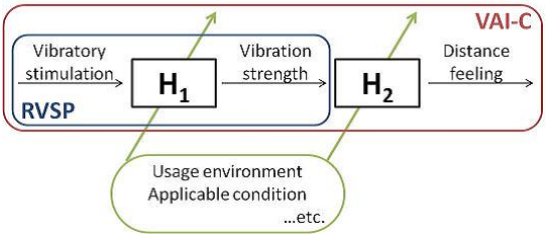


Fig. 6. Hypotheses regarding human perception

Fig. 7. Configuration of VAI-C and participant

4.2 VAI-C Systems

A driving simulation (FORUM8 Corp.) was displayed on three monitors. Participants were three healthy adult men aged 22 to 23 years (average 22.7 years). Participants held the VAI and listened to white noise via headphones, and the experiment was conducted in a darkened room, thus limiting sensory information other than screen information and VAI oscillations. Using the driving simulator

software, we created a virtual straight road on flat ground, with trees on both sides at a fixed interval. Participants were exposed to AVSP with parameter $z = 3$ (Fig. 2). The presentation vibration frequencies varied over 15 settings, from 95 to 165 [kHz] in 5 [kHz] increments, and driving speeds were 30, 60, and 90 [km/h]. This resulted in 45 patterns of varying presentation frequency and driving speed.

4.3 Experimental Exploration of Relations

We performed two experiments to investigate relationships among vibration stimulation, vibration sense, and distance perception. In particular, we wanted to ascertain the following:

(A) Whether differences in driving speed affect vibratory perception.
(B) Whether differences in driving speed affect distance perception.

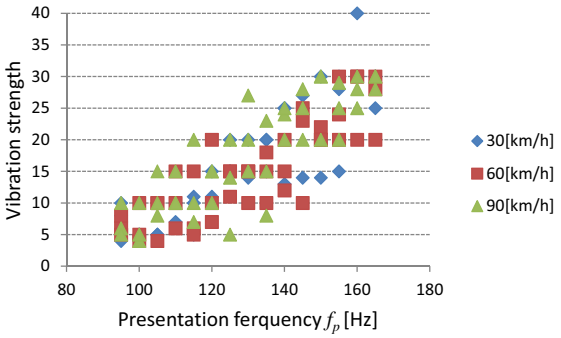
Experimental procedures were as follows:

(A) Driving speed and vibratory sense
  (1) Participants grasped the VAI vibrating at $f_p = 100$ [Hz], and were told to classify perception of this vibration as strength 5.
  (2) Next, they were presented with a new vibration strength, $f_p = 160$ [Hz], which they were told to classify as vibration strength 30.
  (3) Finally, participants were presented with one of the random frequency and speed patterns, and reported the vibration strength.
  (4) Steps (1)–(3) were repeated five times.

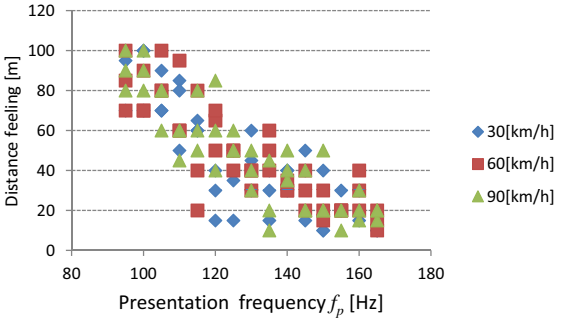
(B) Driving speed and distance perception
  (1) Participants grasped the VAI vibrating at $f_p = 100$ [Hz], and looked at monitors in which a vehicle was 100 [m] ahead. They were told to classify perception of this vibration as that felt for an object 100 [m] away.
  (2) Next, they were presented with a new vibration strength, $f_p = 160$ [Hz], and looked at monitors in which a vehicle was 10 [m] ahead. They were told to classify perception of this vibration as that felt for an object 10 [m] away.
  (3) Finally, participants were presented with one of the random frequency and speed patterns, and reported the distance perception.
  (4) Steps (1)–(3) were repeated five times.

The driving speed was 60 [km/h] in steps (1) and (2). We repeated this process 3 times per day for each participant, with short breaks between presentations. This was repeated over 3 days, resulting in 45 data points regarding vibration strength perception for each participant. Experiment (B) was carried out after completion of experiment (A).

Fig. 8 shows the results for all participants. Both vibration strength and distance perception form a numerical distribution without regard to driving speed, indicating no influence of driving speed on distance or vibration strength perception. This differs from our hypothesis. Our data are currently limited, however, so further investigation is required.



(a) Driving speed and vibration strength.



(b) Driving speed and distance perception.

Fig. 8. The relationship among driving speed, vibration strength, and distance perception

5 Conclusion

This paper showed the repeatability of vibration sensation improvement by the relative vibration sense presentation method. Furthermore, we developed a hypothesis about the relationship of human perception and verified it experimentally.

References

1. Mori, Y., Tanaka, T., Kaneko, S.: Design of Vibration Alert Interface based on Vibration Strength Considering Skin Deformation with Respect to Grip Force. Human Interface Society 12(2), 103–111
2. Yao, H.S., Grant, D., Cruz, M.: Perceived Vibration Strength in Mobile Devices: the Effect of Weight and Frequency. IEEE Transaction on Haptics (2009)

3. Morioka, M., Griffin, M.J.: Magnitude-dependence of equivalent comfort contours for foreandart, lateral and vertical hand-transmitted vibration. Journal of Sound and Vibration 295, 633–648 (2006)
4. Nishiyama, K., Watanabe, S.: Temporary Threshold Shift of Vibratory Sensation After Clasping a Vibrating Handle. International Archives of Occupational and Environmental Health 49, 21–33 (1981)
5. Maeda, S., Griffin, M.J.: Temporary threshold shifts in finger tip vibratory sensation from hand-transmitted vibration and repetitive shock. British Journal of Industrial Medicine 50, 360–367 (1993)
6. Gleason, H.A.: An introduction to descriptive linguistics. Rinehart & Winston, New York, Holt (1961)
7. Minematsu, N., Nishimura, T.: Relative Sense of Speech - A Study of Equality between Speech and Music-. Information Processing Society of Japan (IPSJ). SIG Notes 2005(127), 211–216 (2005)
8. Fechner, G., Adler, H., Howes, D., Boring, E.: Elements of Psychophysics. Rinehart and Winston, New York, Holt (1966)
9. Oyama, T., Imai, S., Wake, T.: New Edition - Handbook for Sense and Perception of Psychology. Seishinsyobo (1994) (in Japanese)

Effect of Light Priming and Encouraging Feedback on the Behavioral and Neural Responses in a General Knowledge Task

Andreea Ioana Sburlea[1], Tsvetomira Tsoneva[1], and Gary Garcia-Molina[2]

Philips Research Europe, The Netherlands
Philips Research North-America, United States
andreea.sburlea@yahoo.com

Abstract. The increase of cognitive demands in society nowadays requires new ways to deal with problems, such as burnout and mental fatigue. Lately, more and more scientifically-based rigorous research in the area of brain-computer interfaces has been done in the quest for restoring and augmenting cognition. The current research work investigates light-based priming and positive reinforcement as possible mediators of cognitive enhancement.

Keywords: priming with light, cognitive enhancement, positive feedback.

1 Introduction

Priming refers to an increased sensitivity to a stimulus due to prior experience. Because priming is believed to occur outside of conscious awareness, it is different from memory that relies on the direct retrieval of information [1]. Priming is an effect of implicit memory. The effects of light-based priming have been widely shown in both humans and animals [2, 3].

Significant research exists on the influence of color on human perception, cognition, and behavior. In [4, 5], blue and green colors are presented as leading to higher cognitive performance than red color, [6, 7] however report the opposite. In [8], it is shown that the red color enhances performance on a detail-oriented task; whereas blue enhances performance on a creative task. These findings together with the ones from [9, 10], suggest that warm colors as being more effective modulators of cognitive performance in a memory related task than cold colors.

The influence of sensory stimuli on cognitive performance in a school context was shown in [11], where exposing underachieving children to olfactory stimulation elicited an increase in performance in a new test by using a scent which was previously associated with high performance in a prior test.

Increased cognitive performance can also result from stereotype priming where people are primed to think about a particular person or profession (the stereotype) exhibiting high cognitive ability, prior to engage in a task requiring cognitive ability. In [12] it is shown that the performance in a general knowledge task of participants

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 744–753, 2013.

primed with the stereotype of a professor is higher than the performance of participants primed with the stereotype of a hooligan.

Feedback and reinforcement can be used in a positive manner to enhance peoples' feelings of competence, which then increases intrinsic motivation. This area, called behavior modification, assumes that behaviors are strengthened when they are rewarded and weakened when they are punished or unrewarded. The stronger the perceived self-efficacy is, the more challenging the goals that people set for themselves become [13].

In a previous study [15], we investigated the influence of light conditioning on cognitive performance. This work can be summarized in three steps: 1) detect (or create) events where a person performs particularly well, 2) apply the targeted light setting with the goal of creating an association between high performance and the light setting, and 3) at a later stage use the light setting to predispose the person for high performance. Three experimental conditions were considered: 1) a control condition, 2) a congruent condition (the association and the test phases had the same light setting) and 3) an incongruent condition (the association and the test phases had different light settings). The cognitive performance associated with each condition was evaluated and positive results were obtained for the congruent condition.

In this study we aim at investigating the behavioral and neural responses as characterized by the electroencephalogram (EEG) of light-based priming and encouraging feedback on a general knowledge cognitive task.

2 Materials and Methods

Twenty healthy volunteers (10 female and 10 male, Mean age = 27.1 and SD = 5.1) participated in the study. All of them had at least a BSc degree. They were randomly assigned to one out of three experimental conditions: a control condition, a congruent-first condition or an incongruent-first condition (see Table 1). All participants signed an informed consent before starting with the experiment. This experiment was approved by the Philips internal ethics commission.

The task of the experiment was a four-choice answer Trivia test which consisted of 4 sets of 25 questions each. There were general knowledge questions belonging to seven different knowledge domains and distributed over three levels of difficulty. All the questions were taken from a Trivia quiz [16]. An example of a question and suggested answers is: "If you suffer from daltonism, you are: a. Color blind, b. Schizophrenic, c. Mute, d. Deaf."

The participants had half a minute to answer to each question. The sets of questions were randomized over the task. EPrimeTM software (from Psychology Software Tools Inc) was used for the presentation of the task [17].

The participants were looking at a 20 inch LCD screen from a distance of 70 cm. Following a short practice session in which no priming was involved, the actual Trivia test started. The light settings (see Fig. 2) were randomly chosen for each participant. After each phase of the experiment the participants were asked to complete a computer-based intrinsic motivation inventory questionnaire [18].

In this study we distinguish three types of feedback which were supposed to modify further performance. True positive feedback – all the questions that were correctly responded received positive feedback. True negative feedback – with a probability of 70%, the questions that were incorrectly responded received negative feedback. Positively biased feedback – incorrectly answered questions could receive positive feedback with a 30% probability. We use the term "encouraging feedback" to refer to the sum of true positive feedback and positively biased feedback.

The feedback was presented on the screen for 3 seconds, starting immediately after the participant's answer or after 30 seconds (time out). For both true positive and positively biased feedback the message displayed on the screen was: "Good job!" followed by the accumulated performance in percentage. For the true negative feedback the message was: "Incorrect answer" followed by the accumulated performance in percentage. In the case of a time-out, the message was: "No response detected".

The study consisted of 4 phases (baseline, association, test1 and test2) each of them lasted for 10 to 15 minutes (see Fig. 1). "P" represents the initial practice phase and "Q" stands for the questionnaire that followed after each phase.

Fig. 1. Timeline of the experiment

The illumination conditions were rendered using 4 Philips LivingColor lamps [20]. The light was projected on the walls. The estimated maximum illumination level was below 100 lux. The colors for this experiment were chosen to be different from each other; complementary colors and not disturbing for the eyes. We distinguish 3 types of light conditions (illumination settings): white light, lightA, lightB.

Fig. 2. Illumination settings (white light, lightA, lightB)

In the baseline phase, the participants performed the task under white light for all conditions. In the association phase, the participants were divided in two groups, one performing the task under lightA and the second under lightB. In the last two phases (testing), each group was further divided into two groups, in which the participants were stimulated with both illumination settings, depending on the condition. Positively biased feedback was given to the participants only during the association phase (see Table 1).

EEG signals were recorded with Biosemi™ Active2 signal acquisition system [19]. The location of the electrodes in the experiment was according to the 10-20 system. Data was recorded from 32 channels (Fp1, AF3, F7, F3, FC1, FC5, T7, C3, CP1, CP5, P7, P3, Pz, PO3, O1, Oz, O2, PO4, P4, P8, CP6, CP2, C4, T8, FC6, FC2, F4, F8,

AF4, Fp2, Fz, Cz). The sampling frequency was 2048 Hz. The EEG signals were first, filtered to remove the 50 Hz power-line noise using a FIR band-stop filter (stop band: 49.9-50.1 Hz). The signals were then subsampled at 256 Hz. Ocular artifact correction was done using the well-known independent component analysis (ICA) based approach [21]. The signals were then re-referenced to the signal average (common average referencing). The resulting signals were band-pass filtered in the 2-25Hz band. This filter permits to attenuate both: DC shifts (low frequencies) and muscle artifacts (high frequencies).

Table 1. The design of the experiment

Phase / Condition	Baseline	Association	Test1	Test2
Control condition	White light, with no positively biased feedback.			
Congruent first condition	White light, with no positively biased feedback.	lightA or lightB positively biased feedback; true positive and true negative feedback.	lightA true positive and true negative feedback lightB	lightB true positive and true negative feedback lightA
Incongruent first condition		lightA or lightB positively biased feedback; true positive and true negative feedback.	lightB true positive and true negative feedback lightA	lightA true positive and true negative feedback lightB

The EEG was segmented in epochs lasting from 500 ms before the onset of the stimulus to 1000 ms after the onset of the stimulus. The EEG signal segments where the energy was not within twice the standard deviation were rejected as they were likely to be movement artifacts.

3 Results and Discussion

The overall performance was measured as a percentage of correct answers. Failing to provide an answer, which happened in 0.2% of the cases, was considered as a wrong answer.

Fig. 3 shows the performance during all the phases, baseline, association, test1 and test2. One should take into account that test1 and test2 are half shared by congruent-first and incongruent-first conditions. On average, the participants answered 47.3% (SD=11.6) of the questions correctly. In the baseline condition the average performance across participants was 48.2% (SD=13.8). In the association condition the average performance was 48.3% (SD=10.8). In the test1 and test 2 conditions, the average performance as 43.7% (SD=7.9) and 48.8% (SD=13.1) respectively (see Fig.3). A t-test revealed no significant difference in the performance between conditions. The third phase presented a smaller variance, having values between 32% and 60% - which might be due to the higher scores in performance displayed on the screen, during association, as a result of the encouraging feedback.

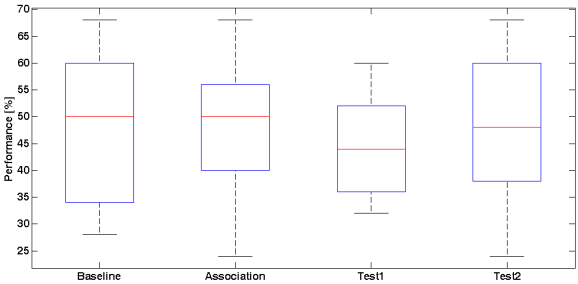


Fig. 3. Overall performance during the experiment

Fig. 4 presents the results of the previous experiment, in which it can be seen that the median performance during the baseline was similar in all the conditions (43%), which means that the participants were equally distributed in terms of proficiency across the three groups. The performance during the association phase of the control condition was very similar to the one of the incongruent condition. The performance in the third phase has lower values in the control (43%) and incongruent conditions (39%), compared to the same phase during congruent condition. The performance during the test phase of the congruent group is higher than the one in the incongruent and control groups, which indicates the effects of light priming. By comparing these results with the control condition, we can also say that, the improvement in the congruent group is caused by both the effects of light and positively biased feedback.

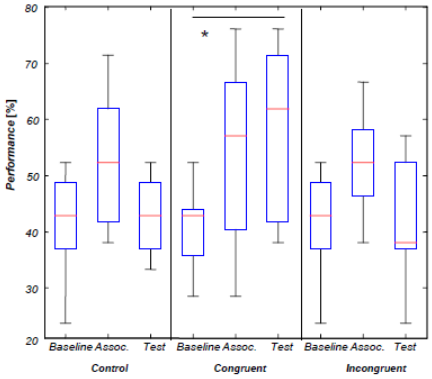


Fig. 4. Overall performance during the first experiment

The average performance level was 48.7% (SD=11.8). The last phase from the control condition had the same average score as the baseline. In the congruent condition, the performance score during association had a larger variance (SD=16.7), average score was 56.3%, higher than in the other conditions. In the congruent condition the test phase had large variance, with an average score of 56% (SD=12.1). During the last phase of the incongruent condition, the average performance score was 39.8% (SD=11.9). The average level of performance decreased as compared to the same phase during the congruent condition.

The participants in the congruent condition had a higher performance than the ones in the incongruent condition. This suggests that the increase is affected by the illumination setting. Furthermore, one of the illumination settings yielded a higher performance improvement, which suggests that the color of light may also play a role.

The experiment design does not permit to distinguish between the effect of light conditioning and that of encouraging feedback. Their combination enhances the performance over all conditions and this may be mediated by an increase in the motivation to perform better.

The analysis of the questionnaire responses yields significant results on the Effort/Importance scale showing that a higher amount of effort was put while performing the last phase of the task in the congruent condition. This also means that performing better during this phase was more important for the participants. The scores of the same scale under control and incongruent conditions had a similar trend, showing that there was no difference in performance in the control and in the incongruent conditions.

To better assess the effect of the intervention (light and encouraging feedback), the last two phases were split according to the corresponding conditions, congruent-first and incongruent-first (see Fig.5).

During the control condition the variance of the performance was large for all phases, except the third one. In the baseline, the average performance was 46% (SD=14). During association, the scores slightly decrease, the average performance is 40.5% (SD=15). During the third phase, the scores decrease even further and the variance of the scores was significantly smaller; the average score was 38% (SD=2.3). In the last phase, most of the participants increased their levels of performance, the average performance was 51% (SD=14.4), which is the highest level of performance over all phases of the control condition.

The congruent-first condition presented a "zig-zag" trend in the levels of performance over phases, with both positive and negative slopes. During the first phase, the average performance score was 45.5% (SD=12.5). Then, the association presented a slight increase in performance, with an average score of 52% (SD=8.8). The third phase, congruent, had the average performance, 46.8% (SD=8) a bit higher than the baseline. The fourth phase presented higher levels of performance, average score was 53% (SD=7.9) of correct answers. There was no significant difference between the phases of the congruent-first condition.

The incongruent-first condition presented a continuous decrease in performance over the phases. The baseline phase presented the average score of performance of 52% (SD=15.7). This is the highest averaged value from all the phases of this condition, but is not significantly higher than the rest. The variance is also very large. During the association, the variance of the performance scores was smaller compared to the baseline, the average score of performance was 48.5% (SD=9.7). The third phase, incongruent, presented the average performance score of 43.5% (SD=8.7). The last phase, congruent, has the largest variance in performance levels, the average level of performance was the same (43.5% (SD=16.1)) as during the previous phase, incongruent, but the variance was larger.

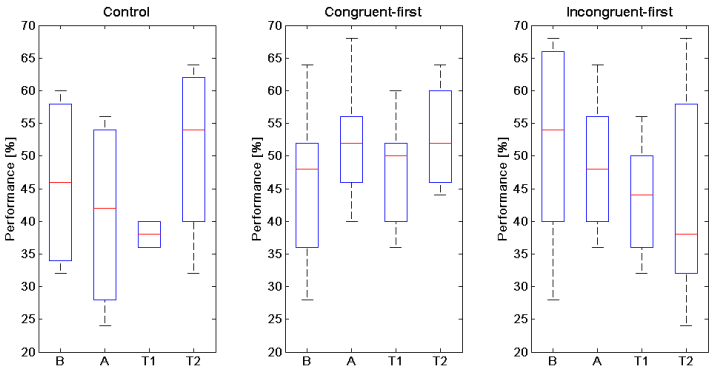


Fig. 5. Box-plots of the performance scores during all the conditions

To compare these results (see Fig. 3) with the ones of the previous experiments (see Fig. 4), the current association phase was split in two parts (see Fig.6).

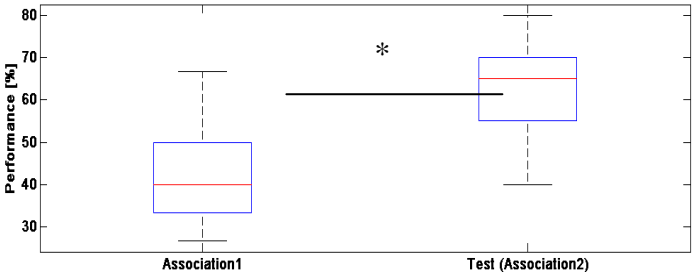


Fig. 6. Splitting the association phase in order to compare the performance results

We assume that in the first part we establish the association and the second part represents the testing. The average performance score of the newly obtained association was 41.7% (SD=11.5), while the new testing phase had a very high the average score of 63.1% (SD=13), compared to the other phase.

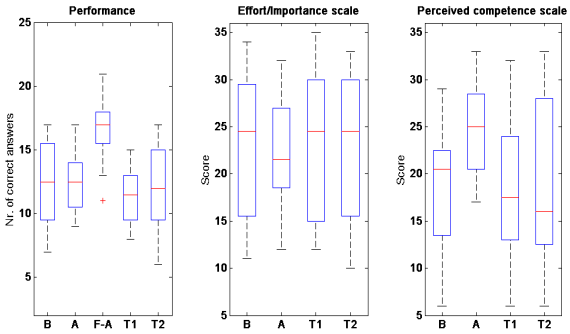


Fig. 7. The number of correct answers in the intervention conditions and the corresponding scores for the Effort/Importance scale and Perceived competence scale

This difference was significant (ttest, t=1, p<0.001) and presented similar trend compared to the congruent condition of the previous experiment. This suggests that a prolonged association is required in order to establish higher performance or that the participant needs to be motivated also during the test phase. Because of the experimental design, we however cannot replicate the incongruent condition of the previous experiment.

The results of the questionnaire together with the performance results during congruent-first and incongruent-first are presented in Fig. 7. The label "F-A" from the performance subplot represents the false (biased) association.

One observation is represented by the correlation between performance and perceived competence. The average performance score over the phases was 50% (SD=11.2), while the positively biased level of performance was 70% (SD=8.9). The Effort/Importance scale presented score in a very large variance and with a similar average value. The Perceived competence scale presented a significant difference between the second and the third phase's scores. A similar decreasing trend was presented in the performance subplot.

Brain activity monitoring is frequently used to gain a better understanding of behavioral results. Here we present the results according to the event related potentials (ERPs) investigation.

Fig. 8 presents the grand average for ERPs 500 milliseconds before the onset of the feedback and 1000 milliseconds after the onset of the feedback. The grand average is presented for the Fz channel. The figure presents 3 different signals corresponding to the type of feedback. The difference between the correctly and incorrectly answered trials Fig. 8 presents a clear peak, around 250-300 milliseconds. This peak is also known as feedback negativity (see [21]). The vertical dashed line at 0 represents the onset of the feedback.

As potential factors that could influence our results, we considered the population knowledge level and the difficulty of the sets. The populations for each experiment have comparable knowledge level. The questions were equally distributed over the sets in terms of difficulty.

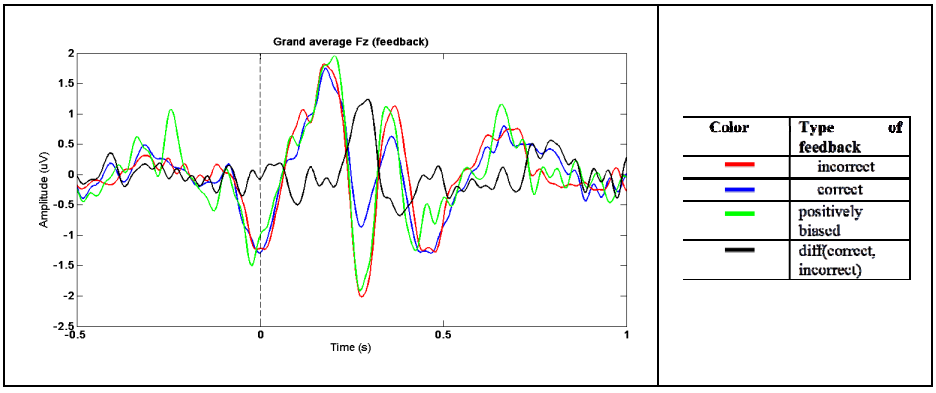


Fig. 8. Grand average across all participants in a [-0.5, 1] second window showing the onset of the feedback for channel Fz

Summarizing all the results that we have so far, brings us to the conclusion that there is no clear influence of light, but the encouraging feedback induces important effects.

The feedback gradually becomes the most salient factor in the process of modulating cognitive performance.

4 Conclusions

The performance of the participants was not strongly influenced by the light intervention. According to the questionnaire results their perceived competence was influenced by positive reinforcement, which played the role of a mediator, leading to a higher performance during that phase. The absence of the encouraging feedback during the next phase led to a decrease in performance and perceived competence.

Regardless of the illumination setting or condition, the feedback seemed to be the most important factor when analyzing the performance scores.

The feedback negativity is a component of the event-related brain potential that is elicited by feedback stimuli associated with unfavorable outcomes. We detected this feature, represented as the difference between correctly and incorrectly answered trials, at 250-300ms after the onset of the feedback. According to the grand average, this feature has the highest magnitude on the frontal cortex, as it is also presented in [14].

The order of the phases had a great impact on performance levels. We observed that regardless of the illumination setting, after association, when the positively biased feedback was introduced, the performance dropped, during the third phase. The order of the congruent phase before or after incongruent, had an impact on performance.

References

1. Schacter, D.L.: Priming and multiple memory systems: Perceptual mechanisms of implicit memory. Journal of Cognitive Neuroscience 4(3), 244–256 (1992)
2. Belcher, M.C., Kluczny, R.: A model for possible effects of light on decision making. *Lighting Design and Application 2, 19–23 (1987)*
3. Knez, I.: Effects of color of light on nonvisual psychological processes. Journal of Environmental Psychology, 201–208 (2001)
4. Kwallek, N., Lewis, C.M.: Appl. Ergon. 21, 275 (1990)
5. Stone, N. J.: J. Environ. Psychol. 23, 63 (2003)
6. Elliot, A.J., Maier, M.A., Moller, A.C., Friedman, R., Meinhardt, J.: J. Exp. Psychol. Gen. 136, 154 (2007)
7. Soldat, S., Sinclair, R.C., Mark, M.M.: Soc. Cogn. 15, 55 (1997)
8. Mehta, R., Zhu, R.: Blue or Red? Exploring the Effect of Color on Cognitive Task Performances. Science 323 (2009)
9. Knez, I.: Effects of color of light on nonvisual psychological processes. Journal of Environmental Psychology, 201–208 (2001)
10. Daggett, W.R., Cobble, J.E., Gertel, S.J.: Color in an optimum learning environment. International Center for Leadership in Education (March 2008)

11. Chu, S.: Olfactory Conditioning of Positive Performance in Humans. Chem. Senses 33, 65–71 (2008)
12. Dijksterhuis, A., van Knippenberg, A.: The relation between perception and behavior, or how to win a game of trivial pursuit. Journal of Personality and Social Psychology 74(4), 865–877 (1998)
13. Bandura, A.: Perceived Self-Efficacy in Cognitive Development and Functioning. Educational Psychology 28(2), 117–148
14. Yeung, N., Holroyd, C.B., Cohen, J.D.: ERP Correlates of Feedback and Reward Processing in the Presence and Absence of Response Choice. Cerebral Cortex 15, 535–544 (2005)
15. Sburlea, A.I., Tsoneva, T., Nijboer, F., Poel, M., Garcia-Molina, G.: Brighten up your mind! Effects of light priming and encouraging feedback on the neural and behavioral responses in a general knowledge task, Master thesis, University of Twente (2012)
16. http://www.quicktrivia.com/
17. Psychology Software Tools Inc., http://www.pstnet.com/
18. http://www.selfdeterminationtheory.org/questionnaires
19. Biosemi, http://www.biosemi.com/
20. http://www.lighting.philips.com/microsite/living_colors/
21. Wallstrom, G.L., et al.: Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods. Elsevier, International Journal of Psychophysiology 53, 105–119 (2004)

Using the Smartphone Accelerometer to Monitor Fall Risk while Playing a Game: The Design and Usability Evaluation of Dance! Don't Fall

Paula Alexandra Silva[1,2], Francisco Nunes[1,3], Ana Vasconcelos[1], Maureen Kerwin[1], Ricardo Moutinho[1], and Pedro Teixeira[1]

[1] Fraunhofer Portugal – AICOS, Rua Alfredo Allen, 455, 4200-135 Porto, Portugal
{ana.vasconcelos,ricardo.moutinho,pedro.teixeira}@fraunhofer.pt, mkkerwin@gmail.com
[2] University of Hawaii, POST Building, 309C, 1680 East-West Rd, Honolulu, HI 96822
paulaalex@hawaii.edu, palexa@gmail.com
[3] Vienna University of Technology, Argentinierstrasse 8, Vienna, Austria
francisco.nunes@igw.tuwien.ac.at

Abstract. Falls are dangerous, and unfortunately common for older adults. Dance! Don't Fall is a game that assesses the quality of the user's locomotion based on data from the accelerometer of a smartphone. By providing a form of exercise, the game may actually reduce fall risk as well as monitoring it. In this paper, we document the development of the prototype and a usability study with ten seniors that suggested the game is well suited to its primary users.

Keywords: Fall risk assessment, older adults, mobile applications, physical activity, dance games.

1 Introduction

Falls are the most common cause of injury and injury-related death among older adults (65+), and one in three older adults suffers a fall every year [1]. To assess fall risk, doctors conduct clinical tests and administer questionnaires [2] [3]. However, these are rarely used before a fall occurs. Furthermore, the infrequency of the tests – once every couple of months – renders them ineffective for detecting sudden changes.

One of the major factors contributing to fall risk is decreased strength and flexibility caused by a lack of physical activity [1], so counteracting the trend of increasingly sedentary lifestyles is a key way to prevent the occurrence of falls.

Our systematic observation and interaction with older adults in a number of senior centers in Portugal has evidenced that older adults particularly enjoy dancing. However, they often are not able to dance because of a lack of specific opportunities, and the difficulty of fitting classes or events into their schedules [4]. Researchers at Fraunhofer Portugal (FhP) - AICOS developed Dance! Don't Fall[1] (DDF), a dance

[1] Dance! Don't Fall is available at http://dancedontfall.projects.fraunhofer.pt

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 754–763, 2013.

game that monitors users' fall risk, while potentially reducing it by promoting systematic exercise. DDF builds upon previous technology developed for conducting and evaluating the gait test using a smartphone as a sensor that the user wears against his or her lower back [5]. This technology was developed with biomedical experts, and includes the clinical gait test and questionnaires in the smartphone. In the same way, DDF provides a means to administer clinical tests at home while it enables and motivates users to exercise regularly, thereby reducing their risk of falling in the first place. The goal of this paper is to present the design and development of DDF as well as the major results of its usability evaluation. Although relevant, assessing the accuracy of the fall algorithm and the long-term efficacy of the game for health purposes are out of the scope of this paper.

2 Related Work

A number of topics contextualize this research, from serious games to games for health, and exergames. Serious games were first approached by Clark Abt, who argued that games should be used as educational tools because of their ability to communicate facts in an efficient way that motivates people to play and, consequently, learn [6]. Later, in the early 2000s, David Rejeski and Ben Sawyer founded the Serious Game Initiative with the goal of spreading the use of games as a means of facing the challenges of the modern world [7]. Since then, serious games have been widely adopted within a variety of areas including military, government, education, business, politics, religion, art, and healthcare [8].

Associated with the Serious Games Initiative, games for health seek to improve healthcare through games that positively impact both mental and physical health [9], educating for healthy habits [11], training and diagnosing cognitive skills [12], complying with rehabilitation programs [14], and improving motor skills [15].

Exergames are a specific kind of game for health that combine exercise and gaming [16]. By adding an element of fun to exercise, exergames can improve seniors' physical and mental health [17]. Exergames have gained popularity with Nintendo's Wii console, but the first commercially successful exergame was Dance Dance Revolution (DDR), which began as an arcade game [18]. DDR players stand on a pad with colored arrows and step on them according to the visual cues on the screen. DDR and similar games are known to provide exercise, helping players become more physically fit and lose weight [19]. However, games like DDR are not adapted to older adults' needs. They include fast-paced music, frequent jumping, and an overload of information on the screen [20]. Moreover, dance pads not only limit the versatility of the stepping pattern, but are also dangerous, as their smooth surfaces can cause them to slip out from underneath players' feet and lead to falls [16]. Nonetheless, DDR inspired researchers to explore dance games for seniors. Smith et al. developed a modified version of DDR and conducted tests with people aged 70+ [21]. Their results showed seniors were able to use the system, but their error rate grew as the step speed and rate increased.

Dancetown is a PC-based exergame specifically designed for seniors. It is similar to DDR as it also uses a dance pad and requires players to follow on-screen cues. Dancetown also includes a rail that can be used with the pad to prevent falls. The

graphics accommodate weakening eyesight, and the game uses music from past generations that may appeal to older users [22]. Studies concluded that the exercise provided by Dancetown is an effective and fun alternative to traditional aerobic exercise; anecdotal findings also indicated that seniors enjoyed playing the game [23].

Finally, DanceAlong [4] is a dance game targeted at seniors that allows players to do "Movioke" – that is, to dance along with scenes from popular movies. This system was tested at a community senior center and players responded enthusiastically, mainly due to the social component of the game.

3 Process

The DDF project followed an iterative and user-centered process with all phases occurring in less than two months. It was the result of an effort of three different teams working concurrently: one designed the user interface, another developed the engine to recognize the dance moves, and another implemented the user interface and developed the communication system that enables multiplayer dances.

The design team, comprised of user interface designers and an element with previous experience in dancing, began by conducting video and live observations of dances in order to identify the characteristics that a dance game system for older adults should have. These revealed that dancing is an activity naturally done together, choreographies are typically simple and repetitive, many dances are derived from traditional dances, and dance steps are in general smooth and small. Furthermore, clapping and producing sounds with the hands seem to be an important part of the dance, not only helping to keep the rhythm, but also stimulating enthusiasm. These were therefore the tenets for the design of DDF, which was initially prototyped on paper and then iteratively refined in terms of functionality, information architecture, and graphic design. In parallel, the design team also chose a song and began working on the dance choreography, which was also iteratively refined, based on feedback from the dance recognition engine team, to ensure the system would be able to detect the dance steps with a sufficient degree of accuracy.

The dance recognition team's work extended the previous work that enabled a smartphone to run a gait test [5]; this algorithm was extended to perceive backward and sideways steps in addition to forward steps. The team developed the rules for step detection by analyzing and testing accelerometer signals from smartphones to discover patterns. The team used Audacity[2] to analyze the music and define the times that steps should occur. They initially implemented the engine in Python using the SciPy open source library of scientific tools, and later ported it to the Android mobile operating system.

[2] Audacity is an open-source software for recording, editing and analysing sound. For more information, refer to: `www.audacity.sourceforge.net`.

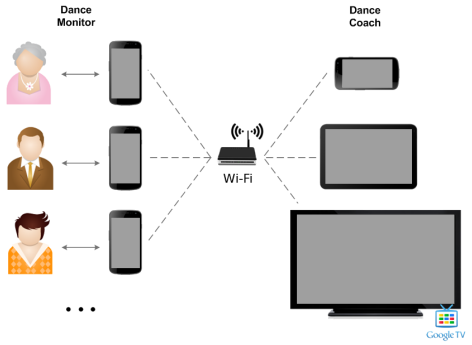


Fig. 1. Multiple devices can be used to play DDF using a Wi-Fi connection

The third team began by implementing the game's user interfaces and basic functionalities. Once the dance engine reached a functional level, it was incorporated into the main Android application and synchronized with the dance choreography for further testing. At this time, the development team focused its resources on the development of the multiplayer component, which was iteratively improved until it was stable and could provide the desired functionality without additional configurations.

4 The DDF System

DDF is a Game for Health that monitors fall risk. To play DDF, the user wears a smartphone on the lower back that tracks his or her dance steps. As users perform choreographed moves along with audio or video dance instructions, the system's algorithms analyze the smartphone's accelerometer data to give feedback on both dance performance and risk of falling. The game gives feedback on four aspects of the dance performance: accuracy, timing, stability, and grooviness. For every dance, the score for each of these factors can be LOW, OK, or HIGH. The fall risk assessment is based on the quality of the user's locomotion and is complemented by a brief questionnaire, presented when a problem appears to exist.

There are three ways to play DDF: Learn, Perform, and Compete. In Learn, a virtual dance coach teaches the individual dance steps and then outlines the choreography. After having learned the choreography, the user may choose Perform to dance alone, or Compete to challenge other friends to a group dance contest. DDF currently features one song and dance, which is based on a simple line dance choreography[3]. As discussed below, more dances should be included in the future.

4.1 Physical Architecture

DDF only requires one smartphone to play the game, but when other Android devices are nearby they can be connected to enhance the user experience and promote social play (Fig. 2). When the game is launched and the device is connected to a Wi-Fi network, the system automatically searches the network for other compatible devices

[3] The team chose a line dance - commonly associated with country-western music and featuring a group of people facing the same direction and performing the same sequence of steps - because the application was intended for presentation at the Mobile Apps Showdown of the Consumer Electronics Show and thus targeted at an American market; furthermore, line dances easily accommodate the characteristics deemed necessary for a dance for seniors.

(Android smartphones or tablets and Google TVs). The user may opt to run additional devices in either dance monitor or dance coach mode, depending on the number and type of devices present.

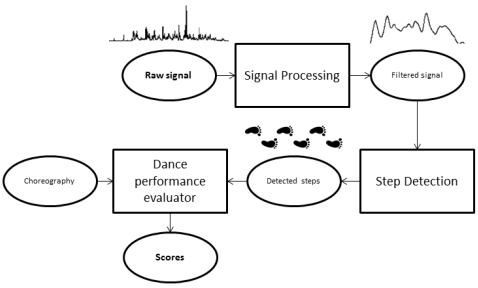


Fig. 2. The processed accelerometer signal is compared to the choreography to score a dance.

Fig. 3. Participants almost exclusively relied on the silhouette to follow the dance.

4.2 Game Modes: Dance Monitor and Dance Coach

The dance monitor is the default mode that includes the game's core mechanics. It requires one smartphone running DDF and is able to detect the steps and play instructions. In addition, the application detects other players in the network and establishes a connection, allowing several players to play simultaneously.

The dance coach mode is an extension that enables players to watch a synchronized dance instructor on another screen. This mode can run on another smartphone, but works best on a tablet or Google TV. Devices running this mode synchronize with the players' smartphones and present the players' ranking at the end of a dance.

4.3 Dance Recognition Engine

The dance monitor mode is powered by the dance recognition engine, which outputs the user's evaluation based on the acceleration of the pelvis, a music track, and a technical definition of the corresponding choreography (comprised of the parts of the dance that are walking-like steps forward, backward, and to the sides). The dance recognition engine contains three modules: a signal processing module, a step detection module, and a dance performance evaluator module. The output of each module respectively serves as the input for the next module (Fig. 3).

The signal-processing module calibrates the raw accelerometer signals, enabling the alignment of the phone's axis with directions relative to the body. Afterwards, a low pass filter with cut-off based on the music frequency is applied, both eliminating the signal's noise and emphasizing the components corresponding to steps. The output of this module is then input to the step detection module.

The step detection module considers different components of the acceleration depending on the direction of movement being performed and outputs a list of perceived steps, each characterized by a timestamp and a direction.

The dance performance evaluator receives a list of steps and compares it to the choreography, which is composed of pairs consisting of a timestamp (the time from the start of the track when the step is supposed to take place) and a direction (the intended movement for the step). Based on the results of this comparison, the module outputs four parameters indicating the performance of the dance:

- Accuracy – the number of correct steps divided by the number of expected steps, a correct step being a step with the same direction as the closest step within a 500 millisecond window in the choreography;
- Timing – the delay between the step's timestamp and the music's time;
- Grooviness – the intuitive sense of dancing in time, evaluated by a combination of pelvic sway and timing; and
- Stability – a measure of how much and how quickly the user performs lateral pelvic displacement.

5 Usability Evaluation of Dance! Don't Fall

The authors conducted a usability study of DDF to identify obstacles to using the system, evaluate the ease of learning and performing the dance, discover users' feelings about the experience, and determine key areas of future research. Ten participants (8 female, 2 male) from senior centers around the city of Porto, Portugal with ages ranging from 60-89 (average 74.2; median 74) took part in the study.

The study took place in the Assisted Living Laboratory at FhP - AICOS. The DDF application ran on both a Google TV and two Android smartphones that the participant and moderator used. The mobile application had been translated into Portuguese, the language used to conduct the tests.

Three team members facilitated the tests. One served as the moderator, giving the introduction and directions, participating in the dance competitions, and generally leading participants through the tests. The other facilitators observed and recorded the participants' behavior and comments; one observer also administered a debriefing interview at the end of the test. One camera recorded the tests from an angle behind the dancers, capturing the display on the TV. Another was attached to the TV and recorded the front view of the dancers, providing a record of the participants' facial expressions and body language.

The test sessions lasted about 45-60 minutes and consisted of an introduction explaining the test and procedures; six tasks, instructed to the participant one at a time; a debriefing interview; and a questionnaire about the participant's health. The tasks required participants to utilize the primary functions of the system, namely: i) input the necessary personal data; ii) accept a dance invitation; iii) comprehend the dance evaluation results; iv) start a dance alone; and v) invite another player to dance. In a normal test situation, each participant performed the dance three times.

6 Findings and Recommendations

Overall, the participants' reaction to the game was positive and they performed well, but the tests did reveal several ways to improve DDF. This section describes the main findings of the usability study.

DDF Is Relevant to the Target Audience. Participants confirmed the two key assumptions behind the game: falls are a frequent issue, and dance is a form of exercise older adults are fond of. Seven had had a fall and feared falling again; the others knew someone who had fallen. All participants stated they liked dancing very much. Eight indicated that they had danced often when younger, while the remaining two said they would like to learn now even though they had not danced much in the past. When inquired about the game itself, nine participants indicated they liked the game very much and would play it at home. The remaining participant did not like the dance style. To address this issue, a future version of the system should offer more variety, namely in terms of styles and levels of difficulty. Besides keeping users interested in the game, offering more difficult dances may also encourage improvement over time and make the game appeal to users with a wider range of fitness levels.

The User Interface Should Be Improved. The participants' ability to learn and perform the dance varied. Several participants performed the dance well from the beginning, some improved markedly with each attempt, while others still could not follow the steps after several attempts. To some extent the variation was caused by differences in physical ability – for instance, three participants turned around so slowly that they fell behind in the dance. But it also seemed to be a matter of the participants' ability to understand DDF's user interface. Nine participants reported they focused solely on the silhouetted demonstrating the dance, ignoring the icons and counter that indicate the current step and the number of times to do it, as well as advising the subsequent step (Fig. 4). This probably indicates an information overload users had when trying to simultaneously interpret the movements of the figure, reproduce them, and attempt to anticipate what to do next. It was not always easy to interpret and mimic the movements of the figure. One participant confused backward and forward steps, not knowing if she was watching the silhouette from the front or the back. Seven participants made errors with the left and right side steps, not sure whether they should mirror the figure or step to their own left or right. Six participants had difficulties performing a step that requires the player to clap and tap their foot at the same time – five only clapped, while one clapped and tapped on opposite beats. Presumably the errors were not due to the difficulty of actually performing the step, but because the participants did not notice the detail of the tapping foot. In addition, to a more clear visual representation of the steps, the inclusion of verbal instructions should be assessed as a way to improve the efficacy in conveying information about the dance steps. This issue, as well as a few others identified in the evaluation (e.g. unclear button labels or problems inputting data), should be addressed in a future version of the game.

DDF Should Emphasize Positive Feedback and Accommodate Beginners. The observers noted that the moderator strayed from the script to encourage and reassure participants: i) after a given participant received low scores (when the moderator typically commented that the participant had done very well considering he or she had just begun learning the dance), and ii) after the participant completed the fall risk questionnaire and received a risk warning (when the moderator assured the participant it was nothing to worry about). Encouraging comments should probably be incorporated in the user interface itself. In total, six participants received low scores and were invited to take the clinical questionnaire. This does not mean that participants had a particularly poor performance, but it shows that the system does not account for the time required to learn the dance. One way to address this is adding the ability to play the dance as a trial that does not receive a score.

7 Discussion

Despite the potential for improvement in the areas discussed above, DDF has several key advantages to other dance games, particularly for older adults. First of all, DDF does not require the purchase of a video game console or physical game media. Anyone who owns a smartphone meeting the minimum requirements can download and play the game at anytime, anywhere. It can be argued that gaming consoles are beginning to make their way into seniors' homes, but our experience with this audience makes us believe that the majority will see them as youth-oriented technology. Smartphones, on the other hand, are becoming more popular and have already conquered the pockets of 22% of U.S. seniors (65+) [24]. For them, DDF significantly lower in not only the effort but also the commitment required to try a new physical activity. Since the smartphone is a multi-purpose device, it has better chances of being welcomed by older adults that believe they are too old to play games. Likewise, DDF accounts for the fact that Google TVs are still fairly uncommon by enabling Android tablets and smartphones to act as the dance coach component in the absence of a Google TV. Furthermore, many of these devices can also duplicate the dance coach display onto a regular television set through HDMI. By supporting connection with large screened devices, DDF transforms itself into a more traditional gaming system and encourages players to dance as a group around the display. While DDF's dances cannot involve movements that are coordinated as pairs – since the system cannot distinguish multiple actors with different sets of movements – this form of group dancing is well-suited to older adults, since not having a partner is one factor that often prevents them from participating in dance activities [4]. Moreover, a study of a projection-based dance system for older adults revealed that participants enjoyed the feeling of dancing with others even though they did not have a designated partner [4].

A key advantage of DDF is the hands-free dance interaction made possible by leveraging the smartphone as a wearable sensor, which provides a more enjoyable and usable experience than games that require the use of an external control device. The most appealing aspect of the new generation of gaming consoles is the use of movement to control the game. However, both Nintendo Wii and PlayStation Move still

use a remote that players hold while playing, limiting the performance of certain gestures. In a study of a digital television exercise application, the participants enjoyed the exercise activities but were distracted by issues related to manipulation of the control device [25]. Likewise, currently available dance games often use game pads that, as stated above, are limiting and possibly dangerous. Microsoft's Kinect and the PlayStation Eye enable users to play without a physical controller. In this vein, DDF players wear the smartphone in a belt, giving them more freedom of movement.

8 Conclusions and Future Work

The usability study produced favorable results, indicating that DDF's objectives align with the goals of the intended primary audience. The evaluation also revealed ways to make the game more effective; but participants successfully completed tasks, enjoyed themselves, and wanted to play again regardless. Overall, the system proved a successful way to utilize the smartphone as a sensor for a dance game as well as assess fall risk through a gait test and questionnaire.

An important subject missing from this evaluation is the question of the game's health aims. How accurately does the system assess the user's dance performance, and how directly does this translate into fall risk? Moreover, what is the impact over time? Do users' results tend to improve as they play more, and does this truly decrease their risk of falling? Such questions are outside the scope of this evaluation but should be addressed in the future, through a more detailed, controlled, long-term study, planned with the collaboration of medical professionals.

References

1. Centers for Disease Control and Prevention. Falls among older adults: an overview, http://www.cdc.gov/HomeandRecreationalSafety/Falls/adultfalls.html
2. Protec-fall.com. Clinical tests evaluating gait and balance, http://www.protec-fall.com/screening-technics/67/clinical-tests-evaluating-gait-and-balance.html
3. Saskatoon Health Region. Fall-risk multi-factor questionnaire, http://www.saskatoonhealthregion.ca/pdf/06_MultiFactor_Falls_Questionnaire.pdf
4. Keyani, P., Hsieh, G., Mutlu, B., Easterday, M., Forlizzi, J.: DanceAlong: supporting positive social exchange and exercise for the elderly through dance. In: CHI 2005, Portland OR, pp. 1541–1544. ACM Press (2005)
5. Guimarães, V., Teixeira, P.M., Monteiro, M.P., Elias, D.: Phone based fall risk prediction. In: Nikita, K.S., Lin, J.C., Fotiadis, D.I., Arredondo Waldmeyer, M.-T. (eds.) MobiHealth 2011. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 83, pp. 135–142. Springer, Heidelberg (2012)
6. Abt, C.C.: Serious Games. Viking Press, New York (1970)
7. Serious Game Initiative, http://tinyurl.com/adh8ekp

8. Michael, D., Chen, S.: Serious Games: Games That Educate, Train, and Inform. Thomson Course Technology, Boston (2006)
9. Games for Health, http://www.gamesforhealth.org/
10. Hoffman, H.G., Patterson, D.R., Carrougher, G.J.: Use of virtual reality for ad-junctive treatment of adult burn pain during physical therapy: a controlled study. Clinical Journal of Pain 16(3), 244–250 (2000)
11. Baranowski, T., Baranowski, J., Cullen, K., Marsh, T., Islam, N., Zakeri, I., Honess-Morreale, L., Demoor, C.: Squire's Quest! Dietary outcome evaluation of a multi-media game. American Journal of Preventive Medicine 24(1), 52–61 (2003)
12. Jimison, H., Pavel, M.: Embedded assessment algorithms within home-based cognitive computer game exercises for elders. In: Proc. of the 28th Annual International Conference of the IEEE, EMBS 2006, pp. 6101–6104. IEEE Xplore, NY (August/September 2006)
13. Rosenberg, D., Depp, C.A., Vahia, I.V., Reichstadt, J., Palmer, B.W., Kerr, J., Norman, G., Jeste, D.V.: Exergames for subsyndromal depression in older adults: a pilot study of novel intervention. Am. J. Geriatr. Psychiatry 18(3), 221–226 (2010)
14. Lange, B., Chien-Yen, C., Suma, E., Newman, B., Rizzo, A.S., Bolas, M.: Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor. In: EMBC 2011, pp. 1831–1834. IEEE Xplore, Boston (August/September 2011)
15. Jung, Y., Koay, J.L., Ng, S.J., Wong, L.C., Kwan, M.L.: Games for a better life: effects of playing Wii games on the well-being of seniors in a long-term care facility. In: IE 2009, Sydney, Australia. ACM Press (December 2009)
16. Bogost, I.: The rhetoric of exergaming (2005), http://bogo.st/cm
17. Brox, E., Luque, L.F., Eversten, G.J., Hernandez, J.E.G.: Exergames for elderly: Social exergames to persuade seniors to increase physical activity. In: PervasiveHealth 2011, Dublin Ireland, pp. 546–549 (May 2011)
18. Arntzen, A.: Game based learning to enhance cognitive and physical capabilities of elderly people: concepts and requirements. World Academy of Science, Engineering and Technology 60, 63–67 (2011)
19. DeMaria, R.: Reset: Changing the Way We Look at Video Games. Berrett-Koehler Publishers, Inc., San Francisco (2007)
20. Lange, B.S., Flynn, S.M., Chang, C.Y., Liang, W., Chieng, C.L., Si, Y., Nanavati, C., Rizzo, A.A.: Development of an interactive stepping game to reduce falls in the elderly. In: ICDVRAT 2010, Valparaíso Chile, pp. 223–228. ICDVRAT (August/September 2010)
21. Smith, S.T., Sherrington, C., Studenski, S., Schoene, D., Lord, S.R.: A novel Dance Dance Revolution (DDR) system for in-home training of stepping ability: basic parameters of system use by older adults. Br. J. Sports Med. 45(5), 441–445 (2011)
22. Canada.com. Senior citizens joining dance revolution, http://tinyurl.com/chh9cvv
23. Thomas, J., Porcari, J., Foster, C., Anders, M.: Dance the calories away: a critical look at Dancetown, an exer-game geared for older adults. ACE Fitness Matters, 9 (November/December 2009)
24. The Nielsen Company. Survey: new U.S. smartphone growth by age and income, http://tinyurl.com/872nlwv
25. Carmichael, A., Rice, M., MacMillan, F., Kirk, A.: Investigating a DTV-based physical activity application to facilitate wellbeing in older adults. In: BCS 2010, Dundee UK, pp. 278–288. ACM Press (September 2010)

Augmented Interaction: Applying the Principles of Augmented Cognition to Human-Technology and Human-Human Interactions

Anna Skinner[1], Lindsay Long[2], Jack Vice[1], John Blitch[3], Cali M. Fidopiastis[4], and Chris Berka[5]

[1] AnthroTronix, Inc., Silver Spring, Maryland
{askinner,jvice} @atinc.com
[2] Clemson University, Clemson, South Carolina
lindsal@clemson.edu
[3] University of Alabama – Birmingham, Birmingham, Alabama
cfidopia@uab.edu
[4] Colorado State University, Fort Collins, Colorado
cfidopia@uab.edu
[5] Advanced Bain Monitoring, Inc., Carlsbad, California
chris@b-alert.com

Abstract. The field of Augmented Cognition (AugCog) has evolved over the past decade from its origins in the Defense Advanced Research Projects Agency (DARPA)-funded research program, emphasizing modulation of closed-loop human-computer interactions within operational environments, to address a broader scope of domains, contexts, and science and technology (S&T) challenges. Among these are challenges related to the underlying theoretical and empirical research questions, as well as the application of advances in the field within contexts such as training and education. This paper summarizes a series of ongoing research and development (R&D) efforts aimed at applying an AugCog-inspired framework to enhance both human-technology and human-human interactions within a variety of training and operational domains.

Keywords: Augmented Cognition, Training, Simulation, Human-Robot Interaction, Adaptive Automation, Neuroscience, Psychophysiological Measures, EEG.

1 Overview

The field of Augmented Cognition (AugCog) has evolved over the past decade from its origins in the Defense Advanced Research Projects Agency (DARPA)-funded Improving Warfighter Information Intake Under Stress (IWIIUS) research program, emphasizing modulation of closed-loop human-computer interactions within operational environments [1, 2], to address a broader scope of domains, contexts, and science and technology (S&T) challenges. Among these are challenges related to the underlying theoretical and empirical research questions, as well as the application of

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 764–773, 2013.

advances in the field within contexts such as training and education. The goal of AugCog is to address the inherent limitations of human operators related to cognitive bottlenecks in information processing such as attention, sensory input, WM, and executive function; and emphasizes real-time monitoring of user cognitive state via behavioral and physiological measures to improve performance through adaptive and augmented human computer interfaces [3]. As data-rich environments become increasingly prevalent, the need for intelligent information management to overcome human information processing limitations is likely to increase within a wide variety of domains such as medicine, education, and information analysis. Additionally, while AugCog methodologies have been applied to domains involving teams of humans, the method of augmentation has primarily been focused on physiologically-based modulation of human-technology interaction [4]. Ongoing research and development (R&D) efforts have begun to emphasize transparency and identical elements in human-robot and human-human interactions, supporting seamless integration of multi-agent human-robot teams, and applying AugCog principles to automated modality selection of information exchange among human team members and between human-robot team members. This paper summarizes a series of ongoing research and development R&D efforts aimed at applying an AugCog-inspired framework to enhance both human-technology and human-human interactions across a variety of training and operational domains.

2 AugCog-Inspired Human-Technology Interaction

Interaction design principles and practices are grounded in both theory and research, guided by academic disciplines such as cognitive psychology and engineering, as well as interdisciplinary fields such as Human Computer Interaction (HCI), human factors, and cognitive ergonomics. The application of AugCog-inspired interaction principles presents a unique paradigmatic shift in the design and use of such products within both training and operational domains.

2.1 AugCog-Inspired Virtual Training Environment Design

Vice, Lathan, Lockerd, & Hitt [5] proposed a novel, AugCog-inspired methodology for determining requirements for virtual environment (VE) design using psychophysiological measures to determine which aspects of VE fidelity and specific VE fidelity configurations would have the highest impact on transfer of training (TOT). Initial validation for this Perceptually-informed Virtual Environment (PerceiVE) design methodology has been demonstrated within a series of empirical studies, indicating that psychophysiological response, and in particular event related potentials (ERPs), may provide a more sensitive index than performance-based measures to changes within underlying cognitive processes occurring during training in VEs, and therefore may be better suited than traditional metrics for highlighting critical fidelity requirements to optimize TOT [6,7]. To better understand the implications for transfer to real world task conditions, Vice, Skinner, Berka, Reinerman-Jones, Barber, Pojman, et al.

[8] compared behavioral and neurological response data between a real world perceptual skills training task and its VE counterpart with varying levels of fidelity. Results indicated that the relationship between physiological response to various VE fidelity configurations and physiological response within an equivalent real world task may be modulated not only by visual feature recognition and processing, but also by higher-order cognitive processes, as evidenced by ERP. Understanding how fidelity variations in VE-based tasks lead to the most efficient processing will inform designers as to which components are responsible for the strongest impartation of skills and ultimately optimize transfer of training [9].

Additional simulation-based training applications that are ripe for exploration using this methodology include remotely piloted aircraft (RPA) training and medical modeling and simulation. Medical simulation-based training reduces risks to human subjects and recues the need for cadaveric and live animal models, supporting training and maintenance of psychomotor skills such as tissue and tool manipulation; cognitive skills related to decision making, declarative and procedural knowledge, and situational awareness; and perceptual skills such as visual feature detection and haptic perception. As in other high-risk training environments, it is the common assumption that a positive linear correlation exists between VE fidelity and skills transfer. However, training on seemingly low fidelity training systems such as the Fundamentals of Laparoscopic Surgery (FLS) video box trainer has repeatedly been demonstrated to translate to complex skills such as interoperative surgical performance, and has become a credentialing criterion for many hospitals [10]. Thus, utilizing the PerceiVE methodology to identify medical simulation design requirements may result in optimized skills instruction, enhancing transfer to real-world medical scenarios.

Skinner, Vice, Berka, & Tan [11] expanded upon this concept, proposing a framework for using psychophysiological measures and feedback within interactive training environments to develop a greater understanding of the processes underlying cross-cultural decision-making and methods for training these critical skills; including detection of variations in information processing and cognitive biases that impact decision-making, interaction within explorable environments, and presentation of relevant cues to facilitate immersion and perspective-taking. This framework suggests that, in particular, neurophysiological metrics such as EEG have the potential to provide an objective measure of cognitive processes involved in attention, perception, and decision-making related to information processing biases; and that eyetracking may support recognition and mitigation of such biases via feedforward and feedback scan patterns, highlighting culturally-relevant perceptual biases. Additionally, this framework incorporates the use of interactive virtual training environments capable of dynamically adapting instruction to individuals based on specific biases exhibited, as well as real-time bias assessment and mitigation.

2.2 AugCog-Inspired Human-Robot Interaction

In addition to simulation-based applications, current research and development is seeking to apply a similar methodology within the context of human-robot interaction (HRI). Woods [12] compared the introduction of automation to a human operated

task to adding another team member who does not necessarily speak the same language and share the same cultural assumptions. Thus, an interface must act as a bridge or translator between humans and automated systems by providing connections and mappings between related concepts in a manner that is partially transparent to the individual human and robot agents. While significant advances are continually made within the domains of robotics and artificial intelligence (AI), the design of robotic control interfaces lacks a validated and scientifically-grounded methodological approach; and currently interface design tends to be an afterthought following development of unmanned systems, with minimal consideration of design and assessment methodologies relying on measures other than those that are purely behaviorally and ergonomically based.

Significant research and development has been invested into physiological sensor-based robot [13] and prosthetic [14, 15] command and control. The application of AugCog principles would expand on this, using physiological signals to measure cognitive states, classify error patterns, and predict cognitive performance degradation within the context of HRI. Vice, Lockerd, and Lathan [16] proposed an AugCog-based approach to multi-modal interface design and implementation, and research conducted under the IWIIUS program demonstrated the use of multi-modal cues and modality switching as a an effective mitigation technique for UAV operations [2]. In recent years, adaptive interfaces have become increasingly prevalent [17], and more specifically, neuroadaptive interfaces are being developed to change in response to meaningful variations in a human user's cognitive and/or emotional states [18]. However, while psychophysiological methods have been investigated in the realm of Adaptive Automation (AA), the vast majority of this work has been oriented on earlier stages of automation (SOA) involved with information acquisition, information analysis, and diagnostic decision support as opposed to the direct action components of unmanned system control (for review see [19]).

Parasuraman, Bahri, Deaton, Morrison, and Barnes [20] identified five primary categories of AA implementation techniques: 1) critical events, 2) operator performance measurement, 3) operator physiological assessment, 4) modeling, and 5) hybrid methods combining one or more of these techniques. Fidopiastis et al. [21] highlight the fact that of these, operator psychophysiological assessment is the only technique that supports unobtrusive real-time operator internal state monitoring without task interruption. Furthermore, this technique may provide the most direct and objective means for assessing and guiding interaction; the dynamic real-time aspects of this methodology preclude the disruptive influence of subjective self-report instruments and secondary task assessments in complex and highly stressful environments while providing temporal resolution on the order of seconds or milliseconds. Byrne and Parasuraman [22] suggest that psychophysiology has two complementary roles within AA research, including assessment of the effects of different forms of automation and the provision of information about the operator that can be integrated with performance measurement and operator modeling to support automation regulation. The advantages posed by the use of non-invasive psychophysiological measurement as a cueing strategy for AA are substantial. While psychophysiological measures may be thought to be most useful for detecting and preventing cognitive

overload, Byrne and Parasuraman [22] have asserted that psychophysiological measures may prove especially useful in the prevention of performance deterioration within underload conditions, which often accompany automation. So-called OOTL (Out Of The Loop) problems have been shown to arise due to human vigilance decrements [23], human complacency [24], and human loss of SA [25]. Thus, as described by Fidopiastis et al. [21], psychophysiologically-based AA has the potential to be applied within high-stress environments in order to alleviate operator workload and fatigue as needed, automating select activities until an operator becomes underloaded and requires additional tasking in order to maintain situation awareness (SA).

The highly structured and quantifiable nature of these measurements also provides a crisp perspective of an operator's cognitive state that can control for individual differences via baseline comparison and minimize the influence of ego and performance bias in risk intensive task / mission sets that require a high degree of confidence as well as technical competence and physical prowess. Fidopiastis et al. [21] highlight the fact that individual differences such as spatial ability and perceived attentional control (PAC) are critical within this context, as demonstrated previously by Chen & Terrence [26].

Parasuraman, Barnes, Cosenzo, and Mulgund [27] specifically demonstrated the effectiveness of AA for supervision of multiple unmanned vehicles. Parasuraman [28] demonstrated the feasibility of matching cardiovascular and cerebral bloodflow-based measures of human mental workload to AA, and more recently, Fidopiastis et al. [21] demonstrated the feasibility of an eye fixation-based workload metric for AA in a simulated robotic control task. Critical to these efforts are the development of reliable measures of cognitive state and performance degradation caused not only by cognitive workload, but also by factors such as fatigue and stress, which may require more sophisticated and sensitive metrics, as well as the integration of various indices. Combining psychophysiological measures with behavioral measures such as validated task battery performance will support the development of hybrid metrics of cognitive function, which amount to more than the sum of their constituent parts, providing more sensitive indices of cognitive state.

Under a current R&D effort our multidisciplinary research team has begun development and validation of a methodology and associated technology tool to support the utilization of multiple, heterogeneous metrics, including operator psychophysiological measures, to drive robotic control interface design and real-time interactions with unmanned systems. This Dynamic Robot Operator Interface Design (DROID) Assessment, Guidance, and Engineering Tool (AGENT) seeks to support instantiation of intelligent sliding autonomy, modality switching, and single versus multi-operator control and feedback offloading. An effective sliding autonomy system should determine the level of individual component autonomy based on maximizing the probability of task or mission accomplishment, taking into account not only operator physical and mental state, but also environmental and task or domain-specific factors. This is critical within the context of military operations in which factors such as rules of engagement, standard operating procedures, and operational tempo may dictate prioritization of tasking and the role of automation, as well as devastating environmental conditions that lie beyond the capabilities and vulnerabilities of human operators.

Finally, as robotic systems increase in complexity beyond anthro-centric limitations to function in such environments, it is equally clear that human cognitive functions are ill suited to do so without the aid of automated modules throughout the spectrum of control.

2.3 Environmental Factors

In addition to incorporating physiological indices of operator state within human-technology interaction design, the context in which such metrics and methodologies are applied must be considered, particularly within military operational environments. A host of factors must be considered beyond cognitive state and information processing limitations, including physical demands on the operator and unique environmental conditions. For example, motion sickness can result from teleoperation tasks, particularly in instances in which the operator is required to teleoperate a robotic asset while in a moving vehicle or on a ship, generating a mismatch between the perceived motion of the unmanned asset and the motion experienced directly by the operator within his or her own environment. AugCog-based systems can be used to gauge the physiological effects of the motion experienced both physically and virtually by the operator. Ideally, a combination of objective and subjective measures could be used to develop validated, multi-dimensional algorithms and constructs to enable effective assessment, prediction, and prevention of motion-induced human performance degradation within a multitude of training and operational environments, including both apparent motion, such as that associated with simulation-based training and teleoperation of remote unmanned vehicles; and actual motion within ground, sea, air, and spaceflight vehicles.

Within the context of naval ship-based operations, ship motion is often a primary contributor to human performance degradation and failures across a wide variety of operational tasks. While motion sickness has been studied extensively, much less research has been dedicated to motion-induced fatigue (e.g., Sopite syndrome symptomology, prevention, effects on performance, and mitigation), and to the complex interactions between motion, fatigue, and stress. Additionally, few studies have explored the constellation of psychophysiological responses associated with motion sickness or the time course of motion sickness, which is non-linear. Neurophysiological metrics have the potential to identify individual differences within a particular task environment, determine metrics that can predict the onset of motion sickness or fatigue, and provide methods for offloading tasks in real-time prior to human performance degradation within the operational environment.

A current effort is being undertaken to design, develop, and validate a Portable Automated Sensor Suite (PASS) Motion-induced User Symptomology Toolkit for Evaluating Readiness (MUSTER) to enable unobtrusive, real-time capture, synchronization, and analysis of environmental, physiological, physical, and subjective measures associated with motion-induced performance degradation within sea-based task environments. This multi-dimensional assessment technology will provide a valuable tool for researchers investigating the effects of motion-induced mishaps, fatigue, and sickness over time, and will also provide a deployable tool for operational use in

determining "fitness for duty". For example, one instantiation of the proposed technology might include a brief set of questions related to motion sickness and fatigue, brief cognitive and psychomotor tests, and the ability to do a rapid physiological sensor reading. Thus, crewmembers could be assessed prior to beginning a shift or prior to conducting high-risk tasks (e.g., on an amphibious vehicle before conducting an amphibious assault) to assess fitness for duty. The embedded algorithms will be developed to flag at-risk individuals, enabling commanding officers to make informed decisions regarding crew shifts and job assignments, and to pull individuals that do not "pass muster" from duty in order to prevent catastrophic performance degradation and errors.

3 AugCog-Inspired Human-Human Interaction

3.1 Cognitive Coupling in Dyads

In addition to modulating human-technology interactions, AugCog principles are beginning to be applied to direct interactions between humans. Stephens, Silbert, and Hasson [29] conducted a groundbreaking experiment in which speaker/listener dyads were monitored simultaneously using functional magnetic resonance imaging (fMRI) to assess neural synchronies between individuals under varied conditions of story comprehension. The results not only provided evidence for detectable spatial and temporal neural coupling in which the listener's brain activity mirrors the speaker's, but also demonstrated that the extent of coupling correlated to the level of story comprehension, and demonstrated that this synchronization ceases under conditions of poor comprehension. During high levels of comprehension, the listeners exhibited predictive anticipatory patterns, with greater the anticipatory speaker–listener coupling corresponding to greater understanding. Stephens and his colleagues argue that this synchronization between production and comprehension-based processes serves as a mechanism by which brains convey information [29], and assert that in many cases the neural processes between brains are coupled, leading to complex synchronized behaviors which must be studied in combination, rather than in isolation in order to be understood [30]. Such brain-to-brain coupling is particularly relevant within the context of dyads in which two humans must collaborate to complete joint tasks, as well as within the context of impartition of knowledge from one individual to another for the purposes of training. Under a current research effort, our team is investigating the use of EEG-based cognitive coupling metrics for an expert/tutor teaching a novice/tutee to complete a complex computer-based task.

3.2 Team Neurodynamics

Recent studies have also shown tremendous promise for the development of EEG-based measures of team cognitive dynamics. For example, Stevens, Galloway, Berka, and Sprang [31] modeled changes in EEG-derived measures of cognitive workload, engagement, and distraction, and explored using neurophysiologic collaboration patterns as an approach for developing a deeper understanding of how teams

collaborate when solving time-critical, complex real-world problems. The resulting cognitive teamwork patterns, termed neural synchronies, were different across six different teams. Stevens, Galloway, Berka, and Behenman [32] suggest that neural synchrony expression may be a reflection of the internal state of team members and of the team as a whole. These studies indicate that non-random patterns of neurophysiologic synchronies can be observed across teams and members of a team when they are engaged in problem solving. This process has been applied to a problem-solving task with students working in teams, as well as navy officers (experts) and officers-in-training (novices) completing a submarine navigation task. Distinct differences were found for expert versus novice neurodynamic synchronies, and novice team neural synchrony metrics were shown to improve (become more like the expert team patterns) over time, providing a potential metric for knowledge and skill acquisition. Furthermore, dynamic detection and classification of individual cognitive states as they relate to team neurocognitive dynamics and performance could be used to identify team members that are not in sync in real time, alerting team leaders to potential underperformance and poor communication in order to support mitigation of team performance degradation via technology-based and interpersonal interventions. This paradigm could be applied across teams of individuals that are both co-located and remotely located, and may in fact provide the most benefit to teams of individuals collaborating over distances in which critical communication elements such as nonverbal cues cannot be relied upon. A vast variety of critical team interaction domains ranging from military operations to surgical teams serve to benefit from such a paradigm.

References

1. St. John, M., Kobus, D.A., Morrison, J.G.: DARPA Augmented Cognition Technical Integration Experiment (TIE). SPAWAR Systems Center Technical Report 1905, San Diego, CA (December 2003)
2. Morrison, J.G., Kobus, D.A., Brown, C.M.: DARPA Improving Warfighter Information Intake Under Stress. Augmented Cognition (2006)
3. Schmorrow, D.D., Kruse, A.A.: Augmented Cognition. In: Bainbridge, W.S. (ed.) Berkshire Encyclopedia of Human-Computer Interaction, pp. 54–59. Berkshire Publishing Group, Great Barrington (2004)
4. Schmorrow, D., Stanney, K., Wilson, G., Young, P.: Augmented Cognition in Human-System Interaction. In: Salvendy, G. (ed.) Handbook of Human Factors & Ergonomics, 3rd edn., pp. 1364–1384. Wiley, Hoboken (2006)
5. Vice, J.M., Lathan, C., Lockerd, A.D., Hitt II, J.M.: Simulation Fidelity Design Informed by Physiologically-based Measurement Tools. In: Proceedings of the 3rd Human Computer Interaction International Conference, Beijing, China, pp. 186–194 (2007)
6. Skinner, A., Vice, J., Lathan, C., Fidopiastis, C.M., Berka, C., Sebrechts, M.: Perceptually-Informed Virtual Environment (PerceiVE) Design Tool. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) Augmented Cognition, HCII 2009. LNCS, vol. 5638, pp. 650–657. Springer, Heidelberg (2009)
7. Skinner, A., Berka, C., Ohara-Long, L., Sebrechts, M.: Impact of Virtual Environment Fidelity on Behavioral and Neurophysiological Response. In: Proceedings of The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2010 (2010)

8. Vice, J., et al.: Use of Neurophysiological Metrics within a Real and Virtual Perceptual Skills Task to Determine Optimal Simulation Fidelity Requirements. In: Shumaker, R. (ed.) Virtual and Mixed Reality, HCII 2011, Part I. LNCS, vol. 6773, pp. 387–399. Springer, Heidelberg (2011)
9. Skinner, A., Sebrechts, M., Fidopiastis, C.M., Berka, C., Vice, J., Lathan, C.: Psychophysiological measures of virtual environment training. In: O'Connor, P.E., Cohn, J.V. (eds.) Human Performance Enhancement in High Risk Environments: Insights, Developments, & Future Directions from Military Research, pp. 129–149. Paeger, Santa Barbara (2010)
10. McCluney, Vassiliou, Kaneva, Cao, Stanbridge, Feldman, Fried: FLS simulator performance predicts intraoperative laparoscopic skill. Surgical Endoscopy 21(11), 1991–1995 (2007)
11. Skinner, A., Vice, J., Berka, C., Tan, V.: Use of psychophysiological measures and interactive environments for understanding and training warfighter cross-cultural decision-making (2012)
12. Woods: Teacher cognition in language teaching: beliefs, decision-making and classroom practice. Cambridge University Press, Cambridge (1996)
13. Millan, J.R., Renkens, F., Mouriño, J., Gerstner, W.: Noninvasive Brain-Actuated Control of a Mobile Robot by Human EEG. IEEE Transactions on Biomedical Engineering 51(6), 1026–1033 (2004)
14. Guger, C., Harkam, W., Hertnaes, C., Pfurtscheller, G.: Prosthetic Control by an EEG-based Brain-Computer Interface (BCI). In: Proc. AAATE 5th European Conference for the Advancement of Assistive Technology, pp. 3–6 (1999)
15. Schwartz, A.B., Cui, X.T., Weber, D.J., Moran, D.W.: Brain-Controlled Interfaces: Movement Restoration with Neural Prosthetics. Neuron 52(1), 205–220 (2006)
16. Vice, J., Lockerd, A., Lathan, C.: Multi-Modal Interfaces for Future Applications of Augmented Cognition. In: Schmorrow, D.D. (ed.) Foundations of Augmented Cognition. Lawrence Erlbaum Associates, Inc. (2005)
17. Haas, M.W., Hettinger, L.J.: Current Research in Adaptive Interfaces. The International Journal of Aviation Psychology 11(2) (2001)
18. Hettinger, L.J., Branco, P., Encarnacao, M., Bonato, P.: Neuroadaptive Technologies: Applying Neuroergonomics to the Design of Advanced Interfaces. Theoretical Issues in Ergonomics Science 4(1-2) (2003)
19. Scerbo, M.: Adaptive Automation. In: Parasuraman, R., Rizzo, M. (eds.) Neuroergonomics: The Brain at Work, pp. 238–252. Oxford University Press, New York (2007)
20. Parasuraman, R., Bahri, T., Deaton, J., Morrison, J., Barnes, M.: Theory and Design of Adaptive Automation in Aviation Systems (Progress Rep. No. NAWCADWAR-92033-60). Naval Air Warfare Center, Warminster (1992)
21. Fidopiastis, C., Drexler, J., Barber, D., Cosenzo, K., Barnes, M., Chen, J., Nicholson, D.: Impact of Automation and Task Load on Unmanned System Operator's Eye Movement Patterns. Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience, 229–238 (2009)
22. Byrne, E.A., Parasuraman, R.: Psychophysiology and Adaptive Automation. Biological Psychology 42(3), 249–268 (1996)
23. Wiener, E.L.: Cockpit Automation. In: Wiener, E.L., Nagel, D.C. (eds.) Human Factors in Aviation, pp. 433–459. Academic Press, San Diego (1988)
24. Parasuraman, R., Molloy, R., Singh, I.L.: Performance Consequences of Automation Induced Complacency. International Journal of Aviation Psychology 3, 1–23 (1993)
25. Endsley, M.R., Kiris, E.O.: The Out-of-the-Loop Performance Problem and Level of Control in Automation. Human Factors 37, 381–394 (1995)

26. Chen, J.Y.C., Terrence, P.I.: Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. Ergonomics 52(8), 907–920 (2009)
27. Parasuraman, R., Barnes, M., Cosenzo, K., Mulgund, S.: Adaptive Automation for Human-Robot Teaming in Future Command and Control Systems. Army Research Lab Aberdeen Proving Ground, MD, Human Research and Engineering Directorate (2007)
28. Parasuraman, R.: Adaptive Automation Matched to Human Mental Workload. NATO Science Series Sub Series I Life and Behavioural Sciences 355, 177–193 (2003)
29. Stephens, G.J., Silbert, L.J., Hasson, U.: Speaker–listener neural coupling underlies successful communication. Proceedings of the National Academy of Sciences 107(32), 14425–14430 (2010)
30. Hasson, U., Ghazanfar, A.A., Galantucci, B., Garrod, S., Keysers, C.: Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends in Cognitive Sciences (2012)
31. Stevens, R.H., Galloway, T., Berka, C., Sprang, M.: Can Neurophysiologic Synchronies Be Detected during Collaborative Teamwork? In: Proceedings: HCI International 2009, San Diego, CA, July 19-24, pp. 271–275 (2009)
32. Stevens, R., Galloway, T., Berka, C., Behenman, A.: Identification and Application of Neurophysiologic Synchronies for Studying the Dynamics of Teamwork. In: Proceedings of the 19th Conference on Behavior Representation in Modeling and Simulation, pp. 21–28 (2010)

Integration of Automated Neural Processing into an Army-Relevant Multitasking Simulation Environment

Jon Touryan[1], Anthony J. Ries[1], Paul Weber[2], and Laurie Gibson[2]

[1] Human Research and Engineering Directorate
U.S. Army Research Laboratory, Aberdeen Proving Ground, MD 21005, USA
{jonathan.o.touryan.ctr,anthony.j.ries2.civ}@mail.mil
[2] Science Applications International Corporation, Louisville, CO, USA
{paul.r.weber,laurie.d.gibson}@saic.com

Abstract. Brain-computer interface technology has experienced a rapid evolution over recent years. Recent studies have demonstrated the feasibility of detecting the presence or absence of targets in visual imagery from the neural response alone. Classification accuracy persists even when the imagery is presented rapidly. While this capability offers significant promise for applications that require humans to process large volumes of imagery, it remains unclear how well this approach will translate to more real-world scenarios. To explore the viability of automated neural processing in an Army-relevant operational context, we designed and built a simulation environment based on a ground vehicle crewstation. Here, we describe the process of integrating and testing the automated neural processing capability within this simulation environment. Our results indicate the potential for significant benefits to be realized by incorporating brain-computer interface technology into future Army systems.

Keywords: Simulator, Brain-Computer Interface (BCI), Visual Search.

1 Introduction

Over the past decade, there has been a substantial improvement in the accuracy of neural signal classification algorithms. One notable area is in the classification of the neural response, as measured via electroencephalography (EEG), elicited by an image. It has been well established that there is a significant difference in the evoked response between images containing a task-relevant target and images without a target [1, 2]. This difference persists even when the image is visible for only a few milliseconds [3]. However, only recently have signal processing algorithms and techniques been sophisticated enough to accurately classify the EEG signal after a single presentation of the image. Initial applications of this technology were for a rapid review or triage of imagery without requiring a manual response to each image [4]. Using a paradigm of rapid image presentation, called rapid serial visual presentation (RSVP), various groups have shown that this approach can identify targets in large ensembles of images an order of magnitude faster than a manual search [5, 6]. However, these

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 774–782, 2013.

initial applications typically used obvious target objects and the only task required of the operator was to focus on the image presentation.

The main goal of this Army-industry collaboration (funded through the Institute of Collaborative Biotechnologies 6.2 Translational Research Program) was to develop a simulation environment to test the performance of state-of-the-art neural classification techniques in a more operational context. The first stage in developing this simulation environment focused on determining the optimal parameters for classification of the neural response in an RSVP paradigm. The parameters investigated included target presentation properties (e.g., size, eccentricity and rate [7]), the effect of changes in attentional state on classification accuracy [8], and the effect of operator multitasking on system performance [9]. The second stage of development, described here, focused on the specific application of the automated neural processing to an Army relevant system. Our intent was to replace the manual visual search task currently utilized to both identify targets and maintain situational awareness in Manned-Ground Vehicles (MGV). Specifically, the RSVP paradigm, in combination with automated classification of the neural response, would replace the manual control of an imaging sensor on the vehicle. Therefore, instead of an operator manipulating the pan-tilt-zoom (PTZ) camera to scan the environment, images of the vehicle's surroundings, containing potential targets, would be rapidly presented and subsequently sorted based on the operator's neural response. The operator could then review the most relevant images for target confirmation.

This second stage of development consisted of two elements. First, we sought to quantify the potential tradeoff of replacing a manual search with RSVP. To accomplish this we conducted an experiment to compare the time-to-target and accuracy of manual search and RSVP paradigms. Second, we developed a simulation environment based on the MGV crewstation. This simulator was designed to switch between the two paradigms and was fully integrated with a real-time EEG processing system. In addition, the simulator incorporated multitasking aspects of the crewstation, including auditory and text communications. Together, these results demonstrate the feasibility and potential benefits of integrating automated neural processing technology into Army systems.

2 Visual Search and RSVP

Sixteen participants were recruited for this experiment, 10 from the general population and six from the project collaborators. They ranged in age from 23 to 55 (mean = 34.8) and included 14 males. Thirteen of the participants were right handed, two were left handed and one was ambidextrous. All individuals recruited from the general population received compensation of $20 per hour. The voluntary, fully informed consent of the persons used in this research was obtained as required by Title 32, Part 219 of the Code of Federal Regulations and Army Regulation 70-25. The investigator has adhered to the policies for the protection of human subjects as prescribed in AR 70-25.

In this experiment, participants alternated between manual visual search and RSVP tasks. In the manual visual search blocks they were required to move a controllable portal (PTZ) over large scenes (1920 x 1080 pixel images) of a simulated urban environment. Images were screen captures from a popular video game (Call of Duty®, Activision Publishing Inc). Above this portal was a low resolution context display of the entire scene with an indicator as to the current location of the portal. The portal was a circular vignette (radius of 150 pixels) initially revealing approximately 3% of the large image; but this could be either increased or decreased based on the zoom factor. Participants were required to use the keyboard (arrow keys, "+", "-") to move the portal and scan the scene for a target (a soldier with a gun). Target identification was indicated by a key press ("t"). Likewise, if no target was found, participants terminated the search with an alternate key press ("n"). The initial placement of the portal was randomly distributed within a given window of 500 pixels around the target. In most cases, the target was not visible in the context display and required the search portal to be detected. The context display served primarily to influence the participant's search path and provide information on likely target locations (doors, windows, cars, etc.).

In the RSVP blocks participants were presented with 100 x 100 pixel image chips representing a region of interest (ROI) from a high resolution image. These ROIs represented salient locations within the image. While the ROIs were manually selected for this experiment, in a real-world application they would be selected by pre-filtering computer vision algorithm [10, 11]. ROIs were displayed at 2 Hz (500 milliseconds), and participants were required to press a button (spacebar) when they saw a target. Since the purpose of the RSVP blocks was to estimate accuracy, only 10 ROIs were included in each RSVP block with a maximum of one target ROI per block. The likelihood of a target was 50 percent for both visual search and RSVP blocks. Participants completed 15 blocks of each task in alternating succession.

A summary of the results for 16 participants is shown in figure 1. As expected, the search time for target-present images was significantly shorter than for target-absent ($p < 0.001$; Wilcoxon rank sum test), 18 seconds versus 60 seconds. While the accuracy for the manual search component was high (mean total accuracy = 0.85), it was substantially lower than for the RSVP component (mean total accuracy = 0.99). There was a significant correlation across participants in accuracy between the search and RSVP components ($r = 0.62$, $p = 0.01$; Pearson's correlation coefficient). For this experiment we decided to keep the RSVP length fixed at 10 ROIs (image chips) or 5 seconds. Under these conditions the RSVP length could be tripled (incorporating up to 30+ ROIs) and still outperform the manual search; requiring less than 15 seconds for completion.

One interesting observation from the manual visual search task was that the search time did not strongly correlate with the initial portal accuracy (i.e., distance from the portal center to the target). The correlation coefficient between search time and initial portal accuracy was 0.09 ($p = 0.07$). Unless the portal was placed within 100 pixels of the target, the initial placement of the portal did not influence the search time because participants followed a search path of potential target locations. This observation

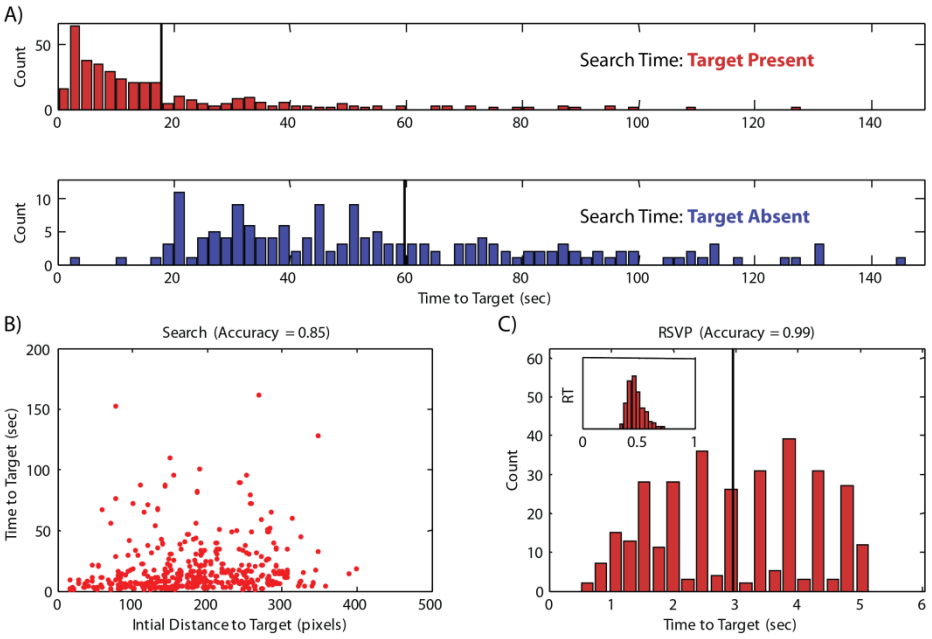


Fig. 1. Manual search and RSVP behavioral summary. A) The manual search time distributions for target-present (red) and target-absent (blue). Vertical lines indicate distribution mean. B) The relationship between initial portal placement and search time. C) The time-to-target in the RSVP condition with a 10 ROI sequence (9 non-target and 1 target image). Inset shows the reaction time (RT) distribution.

speaks directly to the importance of the slew-to-cue accuracy (orientation of the imaging sensor in response to external or environmental cues) in MGV. If the slew-to-cue accuracy can be well quantified, it will be imperative to instruct crewstation operators to stay within the area of initial placement and suppress their instinct to follow a contextual search path. In a similar fashion, the intelligent RSVP should be programmed to give priority to ROIs that fall within the cued area.

Another potential key parameter is portal speed. To test this directly we manipulated portal speed for a subset of participants (N = 10). These participants performed the search experiment in two sessions. In each session their portal speed (in PTZ) was set to a value of either baseline (1x condition) or twice baseline (2x condition). The order of the conditions alternated such that half of the participants had the 1x condition in the first session and half the 2x condition. Over the population we found that there was no significant difference in search time for the two conditions ($p > 0.05$; Wilcoxon rank sum test). However, we did find a significant reduction in search time between session one and two, indicating a significant practice effect ($p < 0.05$). These results suggest that training, rather than PTZ speed, is more important for system performance.

3 Simulator Design

The simulation environment (figure 2) was designed to test the translation of automated neural processing into a more real-world environment. Specifically, this environment was modeled after an Army MGV crewstation. As in the visual search experiment described above, the operator's primary task was to search the environment for targets (soldiers with guns) while the vehicle navigated an urban landscape. The principal performance comparison for this simulator is the speed and accuracy of target detection between a manual search, via a gimbaled camera described above, and an RSVP presentation of pre-filtered ROIs around the vehicle. During the RSVP presentation, each ROI is sorted based on an interest score derived from the evoked brain response [4]. The operator is then immediately presented with the images that generate the highest interest score to manually verify which of the top scoring images contained targets. In addition to this primary task, several other secondary tasks are required of the operator. These secondary tasks are designed to both increase difficulty and replicate real-world environments.

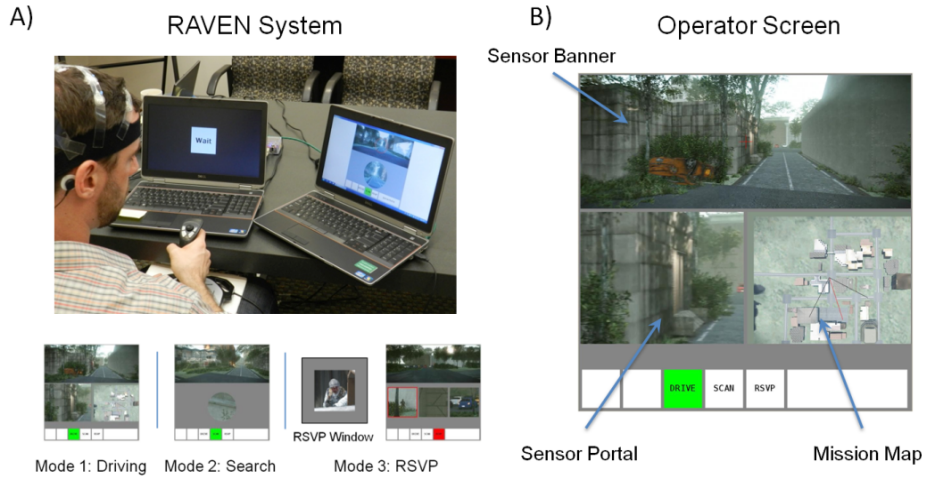


Fig. 2. RSVP-based Adaptive Virtual Environment with Neural-processing (RAVEN) system. A) RAVEN system with main operator screen and RSVP window. The primary task is identification of dismounts while the vehicle is navigating a simulated environment. Secondary tasks include identification of potential IED locations, monitoring and responding to communications (audio and text). The BCI component is engaged during RSVP (Mode 3) and top scoring images are presented on the main operator screen. B) Operator screen during driving (Mode 1) including sensor banner, sensor portal, and mission map windows.

After the optimal design parameters were identified, the simulation environment (called RSVP-based Adaptive Virtual Environment with Neural-processing or RAVEN) was developed to quantify the benefit of incorporating neural processing techniques into this Army relevant operational context. To measure the performance and potential benefits of the RSVP approach, we outlined two validation experiments.

The first experiment, described below, was designed to quantify the speed and accuracy of the two search paradigm with minimal interference from secondary tasks. The second experiment is currently being conducted at the University of California, Santa Barbara, and will focus on the effects of multitasking and task difficulty on overall system performance.

4 Simulator Validation

Fourteen participants were recruited for this experiment, eleven from the general population and three from the project collaborators. Six of these individuals also participated in the visual search experiment (described above). Participants ranged in age from 23 to 59 (mean = 36.9) and included 10 males. Eleven of the participants were right handed and three were left handed. All individuals recruited from the general population received compensation of $20 per hour. The voluntary, fully informed consent of the persons used in this research was obtained as required by Title 32, Part 219 of the Code of Federal Regulations and Army Regulation 70-25. The investigator has adhered to the policies for the protection of human subjects as prescribed in AR 70-25.

In this experiment, participants again alternated between manual visual search and RSVP tasks, this time within the context of the RAVEN simulator (figure 2). Briefly, the task was a simulated patrol of an urban landscape. The vehicle was driven by the computer but the commander (experimental participant) was required to perform several tasks as the vehicle navigated through the environment. The primary task was visual target detection in order to identify threats. At each intersection (24 in all), the vehicle stopped and the participant searched for the target (soldier carrying a gun). At half of the intersections the search was via a controllable portal; in the other half, the search was performed through an RSVP sequence of pre-filtered image chips (ROIs). The majority of intersections contained a target (approximately 80 percent), which could appear at various locations within the scene. A set of potential locations was identified before the experiment but the final target location was randomly chosen by the computer at each intersection. The parameters of the primary task were similar to the visual search experiment described above. However, in this case we used a presentation rate of 5 Hz (200 milliseconds) for the RSVP component. In addition to the intersection search, the participant was required to perform two other tasks while the vehicle was navigating the environment: 1) identify potential IEDs near the roadside (e.g., trash bags, boxes, tires) while the vehicle is moving, and 2) respond to specific radio communications.

Electrophysiological recordings were digitally sampled at 256 Hz from 20 scalp electrodes, located on the standard 10-20 coordinate grid, using an ABM x24 system configured with the single-trial ERP sensor strip (Advanced Brain Monitoring, Carlsbad, CA). EEG was acquired during the entire simulation but real-time analysis, via single trial classification of the evoked response using custom software, was only engaged during RSVP. Individual neural classification models were constructed for each participant from a separate RSVP session prior to the simulator experiment. The

single-trial classification models were linear discriminate functions applied to the neural response elicited by each image. The models were constructed using a machine learning algorithm, described elsewhere [12], and typically achieved area under the ROC curve values between 0.85 and 0.95.

The principal comparison in this validation experiment was speed and accuracy between the manual search and RSVP conditions. At the manual search intersections, participants scanned the environment via the controllable portal until they identified the target or decided that no target was present. At the RSVP intersections, the neural response to each image was scored and the top three ROIs from each intersection search were shown to the participant for final target selection (or confirmation that no target was present). We used metrics of accuracy and time-to-target to quantify the performance in both conditions. Figure 3 shows the results from the 14 participants. At this faster RSVP presentation rate (5 Hz) the difference in mean time-to-target (manual search $\mu = 0.64$ minutes, $\sigma = 0.49$; RSVP $\mu = 0.23$ minutes, $\sigma = 0.11$) was significant ($p < .001$, $t = 7.8$) while the difference in mean accuracy (manual search $\mu = 0.80$, $\sigma = 0.4$; RSVP $\mu = 0.85$, $\sigma = 0.35$) was not statistically significant ($p = 0.259$, $t = -1.13$). These results indicate that even under more realistic conditions, the integration of automated neural processing can enhance overall operational performance.

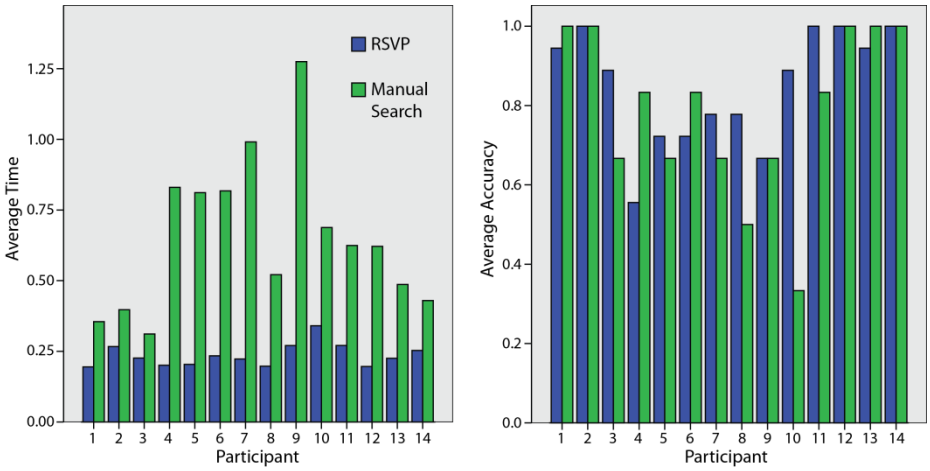


Fig. 3. Performance summary for fourteen participants with 5 Hz RSVP presentation rate. Graphs compare time (in minutes) to find the target and accuracy for RSVP and manual search.

5 Summary

For this study, we defined and implemented a simulation environment that offers a platform to test automated neural processing applications within a real-world context. We identified a common task for MGV operators, manual search of the vehicle environment, which could potentially be replaced with a brain-computer interface. Single-trial classification of the neural response to ROIs of the vehicle surroundings can identify targets more rapidly than the manual search. This performance enhancement

persisted even when the task was embedded within an operation scenario. One important component underling the success of this study is that the accuracy of the single-trial classifier was sufficient to find the target at each intersection, even with a rapid presentation rate (5 Hz). Further tests will be conducted to explore how performance is modulated by increased difficulty and the imposition of multitasking. This simulation environment demonstrates the potential for brain-computer interface technology to meet the challenges of the ever increasing complexity of soldier-systems.

There are several important factors that will influence the success of this brain-computer interface application. One critical factor is how well EEG can be acquired and processed in the real-world vehicle environments. Noise from electrode movement and muscle activation can dramatically affect the EEG signal. Various approaches are being explored to process EEG in real-time within the context of a high noise environment. Another key factor is the performance of the automated systems in the ground vehicle. First, the slew-to-cue accuracy of the external vehicle sensors will directly impact how much benefit can be gained by this neural processing approach. If the automated systems can reliably and accurately locate targets, this triage approach may be unnecessary. Second, the efficacy of the computer vision pre-filtering algorithm will significantly influence the RSVP speed. The length of the RSVP is directly related to how many false alarm ROIs are generated by the pre-filtering algorithm. Given that targets may often be camouflaged or embedded within a dense context, it is likely that numerous false alarms will be detected for every correct target identified. Likewise, the RSVP application described here relies on the pre-filtering algorithm to include the target in the ROI ensemble. Thus, integration of this type of brain-computer application will be a dynamic process that depends on the capabilities of the complementary automated systems.

References

1. Squires, N.K., Squires, K.C., Hillyard, S.A.: Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. Electroencephalography and Clinical Neurophysiology 38, 387–401 (1975)
2. Snyder, E., Hillyard, S.A.: Long-latency evoked potentials to irrelevant, deviant stimuli. Behavioral Biology 16, 319–331 (1976)
3. Thorpe, S., Fize, D., Marlot, C.: Speed of processing in the human visual system. Nature 381, 520–522 (1996)
4. Sajda, P., Pohlmeyer, E., Wang, J., Parra, L.C., Christoforou, C., Dmochowski, J., Hanna, B., Bahlmann, C., Singh, M.K., Chang, S.F.: In a Blink of an Eye and a Switch of a transistor: Cortically Coupled Computer Vision. Proceedings of the IEEE 98, 462–478 (2010)
5. Sajda, P., Gerson, A., Parra, L.: High-throughput image search via single-trial event detection in a rapid serial visual presentation task. In: Proceedings of the First International IEEE EMBS Conference on Neural Engineering, pp. 7–10 (2003)
6. Touryan, J., Gibson, L., Horne, J.H., Weber, P.: Real-time classification of neural signals corresponding to the detection of targets in video imagery. Presented at the International Conference on Applied Human Factors and Ergonomics (July 17, 2010)

7. Gibson, L., Touryan, J., Ries, A., McDowell, K., Cecotti, H., Giesbrecht, B.: Adaptive Integration and Optimization of Automated and Neural Processing Systems-Establishing Neural and Behavioral Benchmarks of Optimized Performance. DTIC Document (2012)
8. Giesbrecht, B., Eckstein, M.P., Abbey, C.K.: Neural decoding of semantic processing during the attentional blink. Journal of Vision. Abstract (2009)
9. Cecotti, H., Kasper, R.W., Elliott, J.C., Eckstein, M.P., Giesbrecht, B.: Multimodal target detection using single trial evoked EEG responses in single and dual-tasks. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, pp. 6311–6314 (2011)
10. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. 40, 1489–1506 (2000)
11. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review 113, 766 (2006)
12. Touryan, J., Gibson, L., Horne, J.H., Weber, P.: Real-time measurement of face recognition in rapid serial visual presentation. Front Psychol. 2, 42 (2011)

Behavioral Biometric Identification on Mobile Devices

Matt Wolff

University of Hawaii, Honolulu, HI 96813, USA
wolffm@hawaii.edu
http://www2.hawaii.edu/~wolffm

Abstract. We show that accelerometers, touch screens and software keyboards, which are standard components of modern mobile phones, can be used to differentiate different test subjects based on the unique interaction characteristics of each subject. This differentiation ability can be applied to authenticate individuals under a continuous authentication scheme. Based on six 15 minute data sets collected from the test subjects utilizing our data collection platform, we extract multiple features from the data and show an ability to accurately identify individuals at a rate of 83 percent using a simple normal distribution of each feature.

Keywords: identification, security.

1 Introduction

Determining the authenticity of the user of a computing device is always an issue when considering a system's security. Standard security practices invoke the use of some type of challenge/response pattern, including passwords, patterns, or biometrics such as a fingerprint. Typically the challenge/response pattern is presented to a user at initial access or at some interval. Initial authentication provides a level of defence at initial access to computing. Continuous authentication, the process in which a user in continually authenticated based on some metrics, can also be applied to authenticate the user during their interaction with a computational device. The use of continuous authentication improves the security of a device by verifying a user after the initial authentication, and continues to do so during the lifetime of user interaction.

Mobile devices are now starting to become the dominant model of human computer interaction. It is estimated that by 2015 over 5 billion devices will be in use worldwide [1]. This growth can be attributed to the low cost, ease of use, and innovative interaction models that are provided by these devices. Many of these mobile devices contain a set of sensors that are capable of tracking low level interaction characteristics of the device user.

In this paper, we look at the different sensors provided by mobile phones, and show that data collected from these sensors can distinguish mobile users by analyzing the user's interaction with the device. Our system observes the interaction characteristics of what we define as behaviormetric data - the subset

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 783–791, 2013.

of biometric data that can be used to express individual behaviors, such as gait, In particular, we collect information from the user/device acceleration, keystrokes, and touch interactions. The analysis engine then extracts key features from the data to support differentiation of users.

2 Related Work

The collection of behaviormetric data is a requirement for conducting any type of behaviormetric analysis on a system capable of collecting such information. Below we take a look at studies that included the collection of behaviormetric data on mobile devices, and the methodologies used in each study.

2.1 Gait-Based Collections

Gait-based classification of users has received recent attention in the field of behaviormetrics, mostly due to the accessibility of mobile devices that contain accelerometer sensors. Boyle et. al. [2] conducted a series of five data collection activities to generate a data set pertaining to sensor readings during the act of walking by test subjects. The authors place a Motorola A855 into the possession of each test subject to collect accelerometer and magnetometer data. In the first three series of data collection activities, the authors collected 33 samples of accelerometer and magnetometer data per experiment on two subjects, with each sample duration being under 60 seconds. The fourth experiment introduced two additional subjects with 28 samples collected per user. In the last experiment, 117 segments of data are collected on each user, with a variance in walking speed across each sample that was not present in previous experiments conducted by the authors.

Mantyjarvi et. al. [3] attached a three-dimensional accelerometer behind the waist of all test subjects. 36 test subjects walked a distance of 20 m at normal, fast and slow speeds. The test was repeated after five days, with 108 total segments of data collected.

Gafurov [4] attached an accelerometer to the leg of 21 study participants, who walked a distance of 35 meters in one direction, and 35 meters back to their original starting position. The data collected was divided into two section, the 35 meters before the turn around, and the 35 meters after the turn around. In a more comprehensive experiment [5], accelerometers were attached to the ankle, hip, pocket, and arm of test subjects while conducting the same walking test as in [4].

In additional gait-based studies, Marc et. al [6] place accelerometers on the ankle of test subjects. 5 subjects in the study walked for 1 minute 8 times a day for five days. Each 1 minute walk introduced different variables, either with different shoes or walking speeds. Over 200 minutes of walking data was collected from the 5 participants. [7] placed an iPhone into the pant pocket of 9 test subject to collect accelerometer and ambient audio data. The 9 participants walked for 2 minutes in indoor and outdoor environments on 3 separate days, with the additional requirement that participants wear different pants on each

day. In [8], 36 subjects placed an Android-based smart phone in the front leg pocket. Subjects were then asked to walk, job, climb up and down stairs for a specified period of time.

2.2 Other Accelerometer Collections

Another popular hand-held device, the television remote was also used for the collection of behaviormetric data. Chang [9] attached an accelerometer to the home remote control of five households. Accelerometer data was collected 24 hours a day for a period of one to three weeks per household.

2.3 Keystroke Collections

Clarke and Furnell [10], focused on the collection of keystroke dynamics based on the entry of telephone numbers and pin codes on a mobile keypad. In their first study, 16 subjects entered 11-digit phone numbers and 4-digit pins into a numerical pad that was typical of a cell phone keyboard input in 2003. In the second study, the authors recruited 30 participants and each participant completed 30 iterations of the entry of 11-digit phone numbers, 4-digit pins, and text messages. The authors collected data on the *inter-keystroke latency* and *hold-time* of the user keystrokes.

In the most comprehensive study of keystroke-dynamics on a mobile device, [11] has twenty-five users participate in a study that collected the press/release time of all keys over a diverse set of Nokia phones running the Symbian operation system over the course of 7 days. Between 2900 and 13713 key hits were recorded per user, correlating to the frequency of keypad use per subject.

[12] recruited 25 users to enter a 4-digit pin into a numeric pad. In this research, in addition to the natural entry method of the test subject, each subject was forced to enter a password with artificial pauses entered into the test entry. Each user created an enrolment set consisting of five password entry recordings, and an additional thirty password entry attempts per entry method. After the test subjects completed the task, they were then asked to pose as *imposters*, where they were given the pin of other test subjects and asked to enter the other subjects pin numbers twice.

[13] recruited forty test subjects, with each subject entering the same 6 password 20 times over four distinct sessions, with each session being a minimum of 10 minutes apart from another session into a alpha-numeric pad. Of note, to enter a character in this type of keyboard can require multiple presses of the same key to select the appropriate character.

[14] combined the use of a number pad with a touch screen to extract hold-time, inter-key duration, finger pressure, and finger position from 10 subjects. Each subject entered the ten digit number thirty times, in consecutive order. Pressure and position were recorded every 20 milliseconds.

3 Data Collection

We developed the software needed to collect readings from the accelerometer, software keyboard, and touch screen of a mobile phone running the Android operation system. Our data collection platform consisted of three major components; an ability to collect raw accelerometer readings across three dimensions, a custom software keyboard that allows for the capturing of keystroke information, and a modification to the android kernel that allows for the collection of touch screen interactions. We used this software to conduct a real-world data collection study of six individual test subjects.

4 Feature Selection

We identified three main areas to extract features from the data set: keystroke dynamics, touch dynamics, and accelerometer signals. For each feature that is selected from the data, we generate a normal distribution of the feature. The normal distribution takes the mean and variance of a feature over the course of the 15 minute test sample, and provides a representation of the probability of a given discrete data point occurring in the test sample. This allows us to compare the same feature across different test subjects to get a reasonable idea regarding the similarity of a feature between two test subjects.

4.1 Keystroke Dynamics

In our analysis of keystroke interaction, we identified several factors that provide a differentiation capability; inter-key duration, key hold time, key-to-key duration, and key press location. Inter-key duration is the measure of the time interval between the release of one key and the press of the next key in the time sequence. Key hold time is a measure of the amount of time a key is held by the user. Key-to-key duration is the inter-key duration between two specific keys. Key press location is the two-dimensional location inside the key where the user initially pressed a key.

4.2 Touch Dynamics

Touch dynamics refers to the extraction of features from the user interaction with a touch screen. In our data collection platform, we are able to determine the location, pressure, and size of each discrete touch event. Typically, multiple discrete touch events are combined to represent a single action. For example, the press of a button may generate a touch event for the down motion, a few events while holding the button down, and a final event when releasing the button.

We separated out touch actions into two distinct groups; taps and gestures. Taps are the collection of touch actions where the distance between the starting and ending point of the touch are below some minimum threshold, and gestures

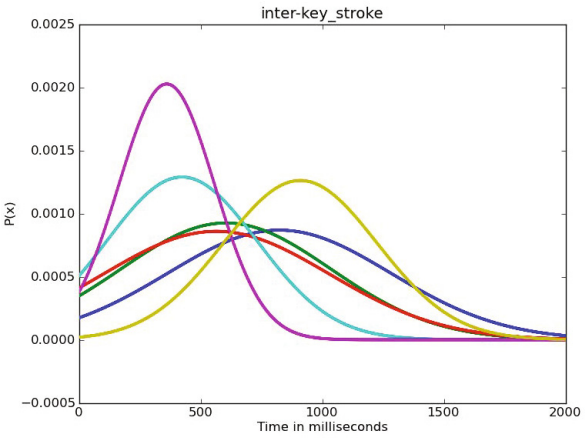


Fig. 1. Normal distribution of each of the six test subject's inter-key duration

consist of the group where the distance is above the same threshold. This separation provides the benefit extracting features from user gestures, such as swipes or scrolling, without being influenced by button presses or selections.

Features of interest identified from the collection of tap activities include the duration of a tap, the two-dimensional location of a tap, the overall pressure of a tap, and the size of a tap. For the collection of gestures, we extracted additional features, inducing the direction of a gesture, the end point, the distance between the start and end of a gesture, the speed of a gesture, and the lateral variance on a gesture. Lateral variance is a measure of the amount of non-direct movement in a gesture. This is calculated by drawing a direct line between the start and end points of a gesture, and calculating the distance between every discrete point that generated the gesture and the direct line between the start and end point.

4.3 Accelerometer Dynamics

Device accelerometers provide acceleration data across three dimensions. In our data collection platform, discrete acceleration events were captured at a rate of about 100 per second during the course of each test subject's interaction. Two main features were extracted from the acceleration data; stability, which is a measure on the variance in acceleration over distinct time periodd, and orientation, which provides an idea of the direction of the x, y, and z axis of the device relative to gravity.

5 User Identification

The main goal of this study is to determine if the behaviormetric data collected from a mobile phone can be used to identify the individual that generated the

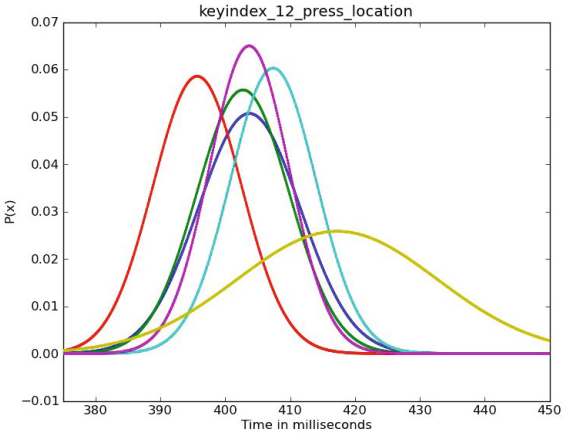


Fig. 2. Normal distribution of the X coordinate of each user's press when hitting key number 12 on the keyboard

data. To test this ability, we remove a random 90 second sample of data from a single user's 15 minutes of data. We then take the random sample, and compare it against the data of each user to find the user most likely to generate the sample.

For each test subject, we calculate a score representing the probability that the 90 second selected sample was generated by the test subject. The score is calculated as follows:

- For each feature (i.e. gesture pressure, z axis acceleration), a continuous normal distribution is generated representing the probability of a discrete event for the test subject and the 90 second sample
- For each feature, we determine the Bhattacharyya coefficient, which represents the amount of overlap between the two statistical samples
- For each feature, the Bhattacharyya coefficient of every test subject is normalized so that the sum of the coefficients for a single feature equal 1
- For each test subject, the normalized coefficient's are summed to generate a score representing the likelihood that the test subject generated the 90 second sample

Figure 5 compares the normal distribution of the 90 second test sample on the x coordinate ending point for a gesture feature to the same feature of each test user. In this figure, the black line represents the normal distribution of the 90 second sample. Based on this figure, one can determine that the purple and blue users have a high probability of having generated the 90 second sample, where as the other four test subjects have a low probability of having generated the 90 second sample.

Overall, for each test subject, we selected six random 90 second samples to test the ability to identify the test subject that generated the data. Using the

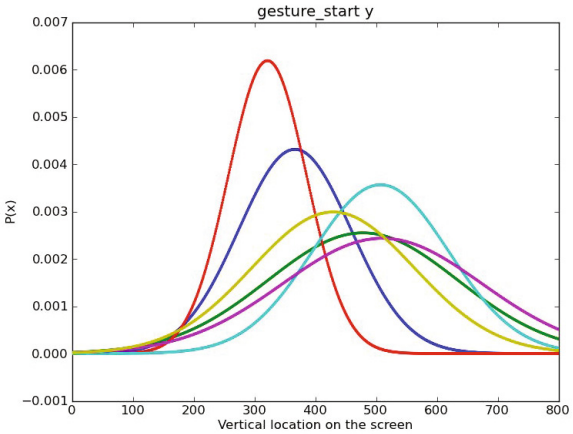


Fig. 3. Normal distribution of the Y coordinate for each user when a user starts a gesture

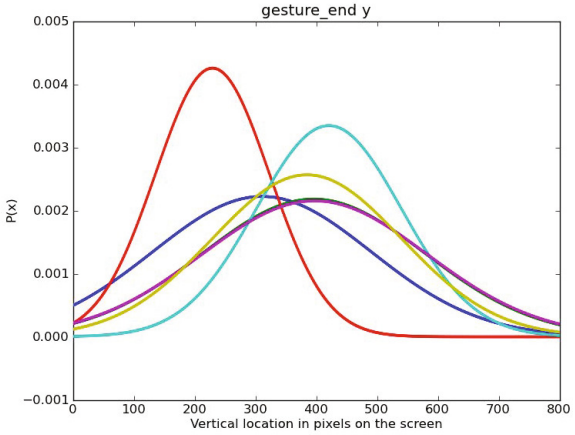


Fig. 4. Normal distribution of the Y coordinate for each user when a user ends a gesture

above technique to compare the selected sample against test subjects, we were able to correctly identify the test subject that generated the 90 second sample 83% of the time. We note that, due to the small sample size, we were not able to include keystroke dynamic features into the identification calculation. Many of the 90 second samples simply did not have enough keystroke information to make any type of determination.

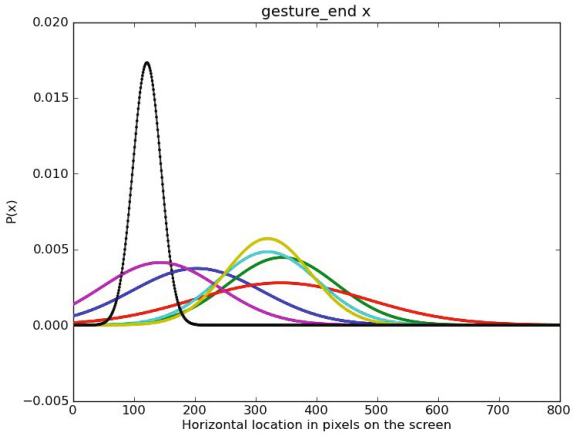


Fig. 5. Black line represents a normal distribution of the X coordinate of the ending point of a gesture of the 90 second sample, compared to all user's normal distribution

6 Conclusion

We have constructed a data collection platform to test the hypothesis that sensors on a mobile device, in particular the accelerometer, touch screen, and keyboard, can be used to differentiate between different users. Based on 15 minutes of real-world device interaction from six test subjects, we were able to correctly identify the test subject that generated a 90 second sample 83% of the time using a subset of features extracted from the data.

Although these findings are encouraging, a larger scale study incorporating more users and larger sample sizes is needed in order to make a more robust determination on the ability to identify users based on their behaviormetric data. In addition, more refined algorithms, as opposed to normal distribution of features as used in this preliminary research, will likely be more effective at user identification. Ultimately, these results provide a simple indication that further study in this area will likely lead to positive results.

References

1. Mcafee, McAfee Threats Report: Second Quarter 2011 (2011)
2. Boyle, M., Klausner, A., Starobinski, D., Trachtenberg, A., Wu, H.: Gait-based User Classification Using Phone Sensors. ipsit.bu.edu, pp. 395–396, `http://ipsit.bu.edu/documents/mobisys11.pdf`
3. Mantyjarvi, J., Lindholm, M., Vildjiounaite, E., Makela, S., Ailisto, H.: Identifying users of portable devices from gait pattern with accelerometers. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings(ICASSP 2005), vol. 2, pp. ii–973. IEEE (2005), `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1415569`

4. Gafurov, D., Helkala, K., Søndrol, T.: Biometric Gait Authentication Using Accelerometer Sensor. Journal of Computers 1(7), 51–59 (2006), http://academypublisher.com/ojs/index.php/jcp/article/view/277
5. Gafurov, D., Snekkenes, E.: Gait Recognition Using Wearable Motion Recording Sensors. EURASIP Journal on Advances in Signal Processing 2009, 1–17 (2009), http://www.hindawi.com/journals/asp/2009/415817.html
6. Bächlin, M., Schumm, J., Roggen, D., Töster, G.: Quantifying gait similarity: User authentication and real-world challenge. Advances in Biometrics, 1040–1049 (2009), http://www.springerlink.com/index/WJ43150286835587.pdf
7. Ketabdar, H., Roshandel, M., Skripko, D.: Towards Implicit Enhancement of Security and User Authentication in Mobile Devices Based on Movement and Audio Analysis. Interactions, 188–191 (2011)
8. Kwapisz, J., Weiss, G., Moore, S.: Cell phone-based biometric identification. In: 2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), pp. 1–7. IEEE (2010), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5634532
9. Chang, K.-h., Hightower, J., Kveton, B.: Inferring identity using accelerometers in television remote controls. In: Tokuda, H., Beigl, M., Friday, A., Brush, A.J.B., Tobe, Y. (eds.) Pervasive 2009. LNCS, vol. 5538, pp. 151–167. Springer, Heidelberg (2009)
10. Furnell, N.L.C.S.M.: Authenticating mobile phone users using keystroke analysis. International Journal of Information Security 6, 1–14 (2007)
11. Zahid, S., Shahzad, M., Khayam, S.A., Farooq, M.: Keystroke-based User Identification on Smart Phones. Technology, 1–18
12. Hwang, S., Cho, S., Park, S.: Keystroke dynamics-based authentication for mobile devices. Computers Security 28(1-2), 85–93 (2009), http://linkinghub.elsevier.com/retrieve/pii/S0167404808000965
13. Maiorana, E., Campisi, P., González-carballo, N., Neri, A.: Keystroke Dynamics Authentication for Mobile Phones. In: SAC, pp. 21–26 (2011)
14. Saevanee, H., Bhattarakosol, P.: Authenticating user using keystroke dynamics and finger pressure. In: Identity, pp. 1–2 (2009)

Author Index