Anna Esposito Antonietta M. Esposito
Alessandro Vinciarelli Rüdiger Hoffmann
Vincent C. Müller (Eds.)

LNCS 7403

Cognitive Behavioural Systems

COST 2102 International Training School
Dresden, Germany, February 2011
Revised Selected Papers

Lecture Notes in Computer Science 7403

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Anna Esposito Antonietta M. Esposito
Alessandro Vinciarelli Rüdiger Hoffmann
Vincent C. Müller (Eds.)

Cognitive Behavioural Systems

COST 2102 International Training School
Dresden, Germany, February 21-26, 2011
Revised Selected Papers

Springer

Volume Editors

Anna Esposito
Seconda Università degli Studi di Napoli
IIASS, Napoli, Italy
E-mail: iiass.annaesp@tin.it

Antonietta M. Esposito
Istituto Nazionale di Geofisica e Vulcanologia
Sezione di Napoli Osservatorio Vesuviano, Napoli, Italy
E-mail: aesposito@ov.ingv.it

Alessandro Vinciarelli
University of Glasgow, School of Computing Science, Glasgow, UK
E-mail: alessandro.vinciarelli@glasgow.ac.uk

Rüdiger Hoffmann
Technische Universität Dresden
Institut für Akustik und Sprachkommunikation, Dresden, Germany
E-mail: ruediger.hoffmann@ias.et.tu-dresden.de

Vincent C. Müller
Anatolia College/ACT
Department of Humanities and Social Sciences, Pylaia, Greece
E-mail: vmueller@act.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-34583-8 e-ISBN 978-3-642-34584-5
DOI 10.1007/978-3-642-34584-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012950318

CR Subject Classification (1998): H.1.2, H.5.5, I.2.7, I.2.9-10, H.5.1-3, H.3.1, H.3.4, I.4.8, I.5.4, G.3, J.4, J.5

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

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)

This book is dedicated to endings and to what is left behind

Preface

This volume brings together the advanced research results obtained by the European COST Action 2102 "Cross Modal Analysis of Verbal and Nonverbal Communication", primarily discussed at the EUCogII-SSPNet-COST2102 International Training School on "Cognitive Behavioural Systems", held in Dresden, Germany, February 21–26 2011 (www.ias.et.tu-dresden.de/ias/cost-2102/).

The school was jointly sponsored by the European Cooperation in Science and Technology (COST, www.cost.eu) in the domain of Information and Communication Technologies (ICT) for disseminating the advances of the research activities developed within the COST Action 2102: "Cross-Modal Analysis of Verbal and Nonverbal Communication" (cost2102.cs.stir.ac.uk); by the European Network of Excellence on Social Signal Processing (SSPNet, www.sspnet.eu) and by the 2nd European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics (EUCogII, www.eucognition.org/).

The main focus of the school was on Cognitive Behavioural Systems. In previous meetings, EUCogII-SSPNet-COST2102 focused on the importance of data processing for gaining enactive knowledge, as well as on the discovery of new processing possibilities that account for new data analysis approaches, coordination of the data flow through synchronization and temporal organization and optimization of the extracted features. The next step will be to discover more natural and intuitive approaches for modelling and uncovering the wealth of information conveyed by humans during interaction for developing realistic and socially believable agents. This moves the research focus to cognitive systems and models of cognitive processes. It has been shown that cognitive processes – such as inference, categorization and memory – are not independent of their physical instantiations. Individual choices, perception and actions emerge and are dynamically affected/enhanced by the interaction between sensory-motor systems and the inhabited environment (including the organizational, cultural and physical context). This interplay carries up instantiations of cognitive behavioural systems.

How can these aspects be modelled in order to bring machine intelligence close to human expectations? Are existing paradigms sufficient or is more research needed on signals and data? How trustful, credible and satisfactory will emotionally-coloured multimodal systems appear to the end user? How will their physical instantiation and appearance affect the human-machine interplay?

The papers accepted in this volume were peer reviewed and include original contributions from early stage researchers. The volume presents new and original research results in the field of human-machine interaction inspired by cognitive behavioural human-human interaction features. The themes covered are cognitive and computational social information processing, emotional and socially believable Human-Computer Interaction (HCI) systems, behavioural and

contextual analysis of interactions, embodiment, perception, linguistics, semantics and sentiment analysis in dialogues and interactions, and algorithmic and computational issues for the automatic recognition and synthesis of emotional states.

The contents have been divided into two scientific sections according to a rough thematic classification. The first section, "Computational Issues in Cognitive Systems", deals with models, algorithms, and heuristic strategies for the recognition and synthesis of behavioural data. The second section, "Behavioural Issues in Cognitive Systems", presents original studies that provide theoretical and behavioural analyses on linguistic and paralinguistic expressions, actions, body movements and activities in human interaction.

The papers included in this book benefited from the lively interactions between the many participants of the successful meeting in Dresden. Over 100 senior and junior researchers gathered for the event.

The editors would like to thank the Management Board of the SSPNet and the ESF COST- ICT Programme for their support in the realization of the school and the publication of this volume. Acknowledgements go in particular to the COST Science Officers, Giseppe Lugano, Matteo Razzanelli, and Aranzazu Sanchez, and the COST 2102 rapporteur, Guntar Balodis, for their constant help, guidance and encouragement. The event owes its success to more individuals than can be named, but notably the members of the Dresden Local Steering Committee, who actively operated for the success of the event. Special appreciation goes to the President of the International Institute for Advanced Scientific Studies (IIASS), and to the Dean and the Director of the Faculty and the Department of Psychology at the Second University of Naples for making available people and resources for editing this volume. The editors are deeply indebted to the contributors for making this book a scientifically stimulating compilation of new and original ideas and to the members of the COST 2102 International Scientific Committee for their rigorous and invaluable scientific revisions, dedication, and priceless selection process.

July 2012

Anna Esposito
Antonietta M. Esposito
Alessandro Vinciarelli
Rüdiger Hoffmann
Vincent C. Müller

Organization

International Steering Committee

Anna Esposito	Second University of Naples and IIASS, Italy
Marcos Faundez-Zanuy	University of Mataro, Barcelona, Spain
Rüdiger Hoffmann	Technische Universtät Dresden (TUD), Germany
Amir Hussain	University of Stirling, UK
Vincent Müller	Anatolia College, Pylaia, Greece
Alessandro Vinciarelli	University of Glasgow, UK

Local Steering Committee

Lutz-Michael Alisch	TUD, Theory of Educational Sciences and Research Methods
Rainer Groh	TUD, Media Design
Rüdiger Hoffmann	TUD, System Theory and Speech Technology
Klaus Kabitzsch	TUD, Technical Information Management Systems
Klaus Meißner	TUD, Multimedia Technology
Boris Velichovsky	TUD, Engineering Psychology and Cognitive Ergonomics
Gerhard Weber	TUD, Human-Computer Interaction

COST 2102 International Scientific Committee

Alberto Abad	INESC-ID Lisboa, Portugal
Samer Al Moubayed	Royal Institute of Technology, Sweden
Uwe Altmann	Friedrich-Schiller-University Jena, Germany
Sigrún María Ammendrup	School of Computer Science, Reykjavik, Iceland
Hicham Atassi	Brno University of Technology, Czech Republic
Nikos Avouris	University of Patras, Greece
Martin Bachwerk	Trinity College, Dublin, Ireland
Ivana Baldasarre	Second University of Naples, Italy
Sandra Baldassarri	Zaragoza University, Spain
Ruth Bahr	University of South Florida, USA
Gérard Bailly	GIPSA-lab, Grenoble, France
Marena Balinova	University of Applied Sciences, Vienna, Austria
Marian Bartlett	University of California, San Diego, USA

Dominik Bauer	RWTH Aachen University, Germany
Sieghard Beller	Universität Freiburg, Germany
Štefan Beňuš	Constantine the Philosopher University, Nitra, Slovak Republic
Niels Ole Bernsen	University of Southern Denmark, Denmark
Jonas Beskow	Royal Institute of Technology, Sweden
Peter Birkholz	RWTH Aachen University, Germany
Horst Bishof	Technical University Graz, Austria
Jean-Francois Bonastre	Université d'Avignon, France
Marek Boháč	Technical University of Liberec, Czech Republic
Elif Bozkurt	Koç University, Istanbul, Turkey
Nikolaos Bourbakis	ITRI, Wright State University, Dayton, USA
Maja Bratanić	University of Zagreb, Croatia
Antonio Calabrese	Istituto di Cibernetica – CNR, Naples, Italy
Erik Cambria	University of Stirling, UK
Paola Campadelli	Università di Milano, Italy
Nick Campbell	University of Dublin, Ireland
Valentín Cardeñoso Payo	Universidad de Valladolid, Spain
Nicoletta Caramelli	Università di Bologna, Italy
Antonio Castro-Fonseca	Universidade de Coimbra, Portugal
Aleksandra Cerekovic	Faculty of Electrical Engineering , Croatia
Peter Cerva	Technical University of Liberec, Czech Republic
Josef Chaloupka	Technical University of Liberec, Czech Republic
Mohamed Chetouani	Université Pierre et Marie Curie, France
Gérard Chollet	CNRS URA-820, ENST, France
Simone Cifani	Università Politecnica delle Marche, Italy
Muzeyyen Ciyiltepe	Gulhane Askeri Tip Academisi, Ankara, Turkey
Anton Cizmar	Technical University of Košice, Slovakia
David Cohen	Université Pierre et Marie Curie, Paris, France
Nicholas Costen	Manchester Metropolitan University, UK
Francesca D'Olimpio	Second University of Naples, Italy
Vlado Delić	University of Novi Sad, Serbia
Céline De Looze	Trinity College, Dublin 2, Ireland
Francesca D'Errico	Università di Roma 3, Italy
Angiola Di Conza	Second University of Naples, Italy
Giuseppe Di Maio	Second University of Naples, Italy
Marion Dohen	ICP, Grenoble, France
Thierry Dutoit	Faculté Polytechnique de Mons, Belgium
Laila Dybkjær	University of Southern Denmark, Denmark
Jens Edlund	Royal Institute of Technology, Sweden
Matthias Eichner	Technische Universität Dresden, Germany

Aly El-Bahrawy	Ain Shams University, Cairo, Egypt
Ciğdem Eroğlu Erdem	Bahçeşehir University, Istanbul, Turkey
Engin Erzin	Koç University, Istanbul, Turkey
Anna Esposito	Second University of Naples, Italy
Antonietta M. Esposito	Osservatorio Vesuviano Napoli, Italy
Joan Fàbregas Peinado	Escola Universitaria de Mataro, Spain
Sascha Fagel	Technische Universität Berlin, Germany
Nikos Fakotakis	University of Patras, Greece
Manuela Farinosi	University of Udine, Italy
Marcos Faúndez-Zanuy	Universidad Politécnica de Cataluña, Spain
Tibor Fegyó	Budapest University of Technology and Economics, Hungary
Fabrizio Ferrara	University of Naples "Federico II", Italy
Dilek Fidan	Ankara University, Turkey
Leopoldina Fortunati	Università di Udine, Italy
Todor Ganchev	University of Patras, Greece
Carmen García-Mateo	University of Vigo, Spain
Vittorio Girotto	Università IUAV di Venezia, Italy
Augusto Gnisci	Second University of Naples, Italy
Milan Gnjatović	University of Novi Sad, Serbia
Bjorn Granstrom	Royal Institute of Technology, Sweden
Marco Grassi	Università Politecnica delle Marche, Italy
Maurice Grinberg	New Bulgarian University, Bulgaria
Jorge Gurlekian	LIS CONICET, Buenos Aires, Argentina
Mohand-Said Hacid	Université Claude Bernard Lyon 1, France
Jaakko Hakulinen	University of Tampere, Finland
Ioannis Hatzilygeroudis	University of Patras, Greece
Immaculada Hernaez	University of the Basque Country, Spain
Javier Hernando	Technical University of Catalonia, Spain
Wolfgang Hess	Universität Bonn, Germany
Dirk Heylen	University of Twente, The Netherlands
Daniel Hládek	Technical University of Košice, Slovak Republic
Rüdiger Hoffmann	Technische Universität Dresden, Germany
Hendri Hondorp	University of Twente, The Netherlands
David House	Royal Institute of Technology, Sweden
Evgenia Hristova	New Bulgarian University, Sofia, Bulgaria
Stephan Hübler	Dresden University of Technology, Gremany
Isabelle Hupont	Aragon Institute of Technology, Zaragoza, Spain
Amir Hussain	University of Stirling, UK
Viktor Imre	Budapest University of Technology and Economics, Hungary

Ewa Jarmolowicz	Adam Mickiewicz University, Poznan, Poland
Kristiina Jokinen	University of Helsinki, Finland
Jozef Juhár	Technical University Košice, Slovak Republic
Zdravko Kacic	University of Maribor, Slovenia
Bridget Kane	Trinity College Dublin, Ireland
Jim Kannampuzha	RWTH Aachen University, Germany
Maciej Karpinski	Adam Mickiewicz University, Poznan, Poland
Eric Keller	Université de Lausanne, Switzerland
Adam Kendon	University of Pennsylvania, USA
Stefan Kopp	University of Bielefeld, Germany
Jacques Koreman	University of Science and Technology, Norway
Theodoros Kostoulas	University of Patras, Greece
Maria Koutsombogera	Inst. for Language and Speech Processing, Greece
Robert Krauss	Columbia University, New York, USA
Bernd Kröger	RWTH Aachen University, Germany
Gernot Kubin	Graz University of Technology, Austria
Olga Kulyk	University of Twente, The Netherlands
Alida Labella	Second University of Naples, Italy
Emilian Lalev	New Bulgarian University, Bulgaria
Yiannis Laouris	Cyprus Neuroscience and Technology Institute, Cyprus
Anne-Maria Laukkanen	University of Tampere, Finland
Amélie Lelong	GIPSA-lab, Grenoble, France
Borge Lindberg	Aalborg University, Denmark
Saturnino Luz	Trinity College Dublin, Ireland
Wojciech Majewski	Wroclaw University of Technology, Poland
Pantelis Makris	Neuroscience and Technology Institute, Cyprus
Kenneth Manktelow	University of Wolverhampton, UK
Raffaele Martone	Second University of Naples, Italy
Rytis Maskeliunas	Kaunas University of Technology, Lithuania
Dominic Massaro	University of California, Santa Cruz, USA
Olimpia Matarazzo	Second University of Naples, Italy
Christoph Mayer	Technische Universität München, Germany
David McNeill	University of Chicago, USA
Jiří Mekyska	Brno University of Technology, Czech Republic
Nicola Melone	Second University of Naples, Italy
Katya Mihaylova	University of National and World Economy, Sofia, Bulgaria
Péter Mihajlik	Budapest University of Technology and Economics, Hungary
Michal Mirilovič	Technical University of Košice, Slovak Republic

Izidor Mlakar	Roboti c.s. d.o.o, Maribor, Slovenia
Helena Moniz	INESC-ID, Lisboa, Portugal,
Tamás Mozsolics	Budapest University of Technology and Economics, Hungary
Vincent C. Müller	Anatolia College/ACT, Pylaia, Greece
Peter Murphy	University of Limerick, Limerick, Ireland
Antonio Natale	University of Salerno and IIASS, Italy
Costanza Navarretta	University of Copenhagen, Denmark
Eva Navas	Escuela Superior de Ingenieros, Bilbao, Spain
Delroy Nelson	University College London, UK
Géza Németh	University of Technology and Economics, Budapest, Hungary
Friedrich Neubarth	Austrian Research Inst. Artificial Intelligence, Austria
Christiane Neuschaefer-Rube	RWTH Aachen University, Germany
Giovanna Nigro	Second University of Naples, Italy
Anton Nijholt	Universiteit Twente, The Netherlands
Jan Nouza	Technical University of Liberec, Czech Republic
Michele Nucci	Università Politecnica delle Marche, Italy
Catharine Oertel	Trinity College Dublin, Ireland
Stanislav Ondáš	Technical University of Košice, Slovak Republic
Rieks Op den Akker	University of Twente , The Netherlands
Karel Paleček	Technical University of Liberec, Czech Republic
Igor Pandzic	Faculty of Electrical Engineering, Croatia
Harris Papageorgiou	Institute for Language and Speech Processing, Greece
Kinga Papay	University of Debrecen, Hungary
Paolo Parmeggiani	Università degli Studi di Udine, Italy
Ana Pavia	Spoken Language Systems Laboratory, Lisbon, Portugal
Paolo Pedone	Second University of Naples, Italy
Tomislav Pejsa	University of Zagreb, Croatia
Catherine Pelachaud	CNRS, Télécom ParisTech, France
Bojan Petek	University of Ljubljana, Slovenia
Harmut R. Pfitzinger	University of Munich, Germany
Francesco Piazza	Università degli Studi di Ancona, Italy
Neda Pintaric	University of Zagreb, Croatia
Matúš Pleva	Technical University of Košice, Slovak Republic
Isabella Poggi	Università di Roma 3, Italy
Guy Politzer	Université de Paris VIII, France
Jan Prazak	Technical University of Liberec, Czech Republic
Ken Prepin	Télécom ParisTech, France
Jiří Přibil	Academy of Sciences, Czech Republic
Anna Přibilová	Slovak University of Technology, Slovak Republic

Emanuele Principi	Università Politecnica delle Marche, Italy
Michael Pucher	Telecommunications Research Center Vienna, Austria
Jurate Puniene	Kaunas University of Technology, Lithuania
Ana Cristina Quelhas	Instituto Superior de Psicologia Aplicada, Lisbon, Portugal
Kari-Jouko Räihä	University of Tampere, Finland
Roxanne Raine	University of Twente, The Netherlands
Giuliana Ramella	Istituto di Cibernetica – CNR, Naples, Italy
Fabian Ramseyer	University Hospital of Psychiatry, Bern, Switzerland
José Rebelo	Universidade de Coimbra, Portugal
Peter Reichl	FTW Telecommunications Research Center, Austria
Luigi Maria Ricciardi	Università di Napoli "Federico II", Italy
Maria Teresa Riviello	Second University of Naples and IIASS, Italy
Matej Rojc	University of Maribor, Slovenia
Nicla Rossini	Università del Piemonte Orientale, Italy
Rudi Rotili	Università Politecnica delle Marche, Italy
Algimantas Rudzionis	Kaunas University of Technology, Lithuania
Vytautas Rudzionis	Kaunas University of Technology, Lithuania
Hugo L. Rufiner	Universidad Nacional de Entre Ríos, Argentina
Milan Rusko	Slovak Academy of Sciences, Slovak Republic
Zsófia Ruttkay	Pazmany Peter Catholic University, Hungary
Yoshinori Sagisaka	Waseda University, Tokyo, Japan
Bartolomeo Sapio	Fondazione Ugo Bordoni, Rome, Italy
Mauro Sarrica	University of Padua, Italy
Gellért Sárosi	Budapest University of Technology and Economics, Hungary
Gaetano Scarpetta	University of Salerno and IIASS, Italy
Silvia Scarpetta	Salerno University, Italy
Stefan Scherer	Ulm University, Germany
Ralph Schnitker	Aachen University, Germany
Jean Schoentgen	Université Libre de Bruxelles, Belgium
Björn Schuller	Technische Universität München, Germany
Milan Sečujski	University of Novi Sad, Serbia
Stefanie Shattuck-Hufnagel	MIT, Research Laboratory of Electronics, USA
Marcin Skowron	Austrian Research Inst. for Art. Intelligence, Austria
Jan Silovsky	Technical University of Liberec, Czech Republic
Zdeněk Smékal	Brno University of Technology, Czech Republic
Stefano Squartini	Università Politecnica delle Marche, Italy
Piotr Staroniewicz	Wroclaw University of Technology, Poland
Ján Staš	Technical University of Košice, Slovakia
Vojtěch Stejskal	Brno University of Technology, Czech Republic

Marian Stewart-Bartlett	University of California, San Diego, USA
Xiaofan Sun	University of Twente, The Netherlands
Jing Su	Trinity College Dublin, Ireland
Dávid Sztahó	Budapest University of Technology and Economics, Hungary
Jianhua Tao	Chinese Academy of Sciences, P. R. China
Balázs Tarján	Budapest University of Technology and Economics, Hungary
Jure F. Tasič	University of Ljubljana, Slovenia
Murat Tekalp	Koc University, Istanbul, Turkey
Kristinn Thórisson	Reykjavík University, Iceland
Isabel Trancoso	Spoken Language Systems Laboratory, Portugal
Luigi Trojano	Second University of Naples, Italy
Wolfgang Tschacher	University of Bern, Switzerland
Markku Turunen	University of Tampere, Finland
Henk Van den Heuvel	Radboud University Nijmegen, The Netherlands
Betsy van Dijk	University of Twente, The Netherlands
Giovanni Vecchiato	Università "La Sapienza", Roma, Italy
Leticia Vicente-Rasoamalala	Alchi Prefectural Univesity, Japan
Robert Vich	Academy of Sciences, Czech Republic
Klára Vicsi	Budapest University, Hungary
Hannes Högni Vilhjálmsson	Reykjavík University, Iceland
Jane Vincent	University of Surrey, Guilford, UK
Alessandro Vinciarelli	University of Glasgow, UK
Laura Vincze	Università di Roma 3, Italy
Carl Vogel	Trinity College, Dublin, Ireland
Jan Volín	Charles University, Czech Republic
Rosa Volpe	Université de Perpignan, France
Martin Vondra	Academy of Sciences, Czech Republic
Pascal Wagner-Egger	Fribourg University, Switzerland
Yorick Wilks	University of Sheffield, UK
Matthias Wimmer	Institute for Informatics, Munich, Germany
Matthias Wolf	Technische Universität Dresden, Germany
Bencie Woll	University College London, UK
Bayya Yegnanarayana	International Institute of Information Technology, India
Vanda Lucia Zammuner	University of Padua, Italy
Jerneja Žganec Gros	Alpineon, Development and Research, Slovenia
Goranka Zoric	Faculty of Electrical Engineering, Croatia

Sponsors

The following organizations sponsored and supported the International Conference

- European COST Action 2102 "*Cross-Modal Analysis of Verbal and Nonverbal Communication*" (cost2102.cs.stir.ac.uk)


ESF Provide the COST Office through and EC contract


COST is supported by the EU RTD Framework programme

COST- the acronym for European Cooperation in Science and Technology- is the oldest and widest European intergovernmental network for cooperation in research. Established by the Ministerial Conference in November 1971, COST is presently used by the scientific communities of 36 European countries to cooperate in common research projects supported by national funds.

The funds provided by COST - less than 1% of the total value of the projects - support the COST cooperation networks (COST Actions) through which, with EUR 30 million per year, more than 30 000 European scientists are involved in research having a total value which exceeds EUR 2 billion per year. This is the financial worth of the European added value which COST achieves.

A "bottom up approach" (the initiative of launching a COST Action comes from the European scientists themselves), "à la carte participation" (only countries interested in the Action participate), "equality of access" (participation is open also to the scientific communities of countries not belonging to the European Union) and "flexible structure" (easy implementation and light management of the research initiatives) are the main characteristics of COST.

As precursor of advanced multidisciplinary research COST has a very important role for the realisation of the European Research Area (ERA) anticipating and complementing the activities of the Framework Programmes, constituting a "bridge" towards the scientific communities of emerging countries, increasing the mobility of researchers across Europe and fostering the establishment of "Networks of Excellence" in many key scientific domains such as: Biomedicine and Molecular Biosciences; Food and Agriculture; Forests, their Products and Services; Materials, Physical and Nanosciences; Chemistry and Molecular Sciences and Technologies; Earth System Science

and Environmental Management; Information and Communication Technologies; Transport and Urban Development; Individuals, Societies Cultures and Health. It covers basic and more applied research and also addresses issues of pre-normative nature or of societal importance.
Web: http://www.cost.eu

- SSPnet: European Network on Social Signal Processing, http://sspnet.eu/

The ability to understand and manage social signals of a person we are communicating with is the core of social intelligence. Social intelligence is a facet of human intelligence that has been argued to be indispensable and perhaps the most important for success in life. Although each one of us understands the importance of social signals in everyday life situations, and in spite of recent advances in machine analysis and synthesis of relevant behavioral cues like blinks, smiles, crossed arms, head nods, laughter, etc., the research efforts in machine analysis and synthesis of human social signals like empathy, politeness, and (dis)agreement, are few and tentative. The main reasons for this are the absence of a research agenda and the lack of suitable resources for experimentation.

The mission of the SSPNet is to create a sufficient momentum by integrating an existing large amount of knowledge and available resources in Social Signal Processing (SSP) research domains including cognitive modeling, machine understanding, and synthesizing social behavior, and so:

- Enable the creation of the European and world research agenda in SSP;
- Provide efficient and effective access to SSP-relevant tools and data repositories to the research community within and beyond the SSPNet, and
- Further develop complementary and multidisciplinary expertise necessary for pushing forward the cutting edge of the research in SSP.

The collective SSPNet research effort is directed towards integration of existing SSP theories and technologies, and towards identification and exploration of potentials and limitations in SSP. More specifically, the framework of the SSPNet will revolve around two research foci selected for their primacy and significance: Human-Human Interaction (HHI) and Human-Computer Interaction (HCI). A particular scientific challenge that binds the SSPNet partners is the synergetic combination of human-human interaction models, and automated tools for human behavior sensing and synthesis, within socially-adept multimodal interfaces.

- **EUCogII: 2nd European Network for the** ***Advancement of Artificial Cognitive Systems, Interaction and Robotics*** **(http://www.eucognition.org/**

- **School of Computing Science, University of Glaskow, Scotland, UK**
- **Department of Psychology, Second University of Naples, Caserta, Italy**
- **Technische Universität Dresden, Institut für Akustik und Sprachkommunikation, Dresden, Germany**
- **International Institute for Advanced Scientific Studies** "E.R. Caianiello" **IIASS,** www.iiassvietri.it/
- **Società Italiana Reti Neuroniche, SIREN, www.associazionesiren.org/**
- **Regione Campania** and **Provincia di Salerno, Italy**

Table of Contents

Computational Issues in Cognitive Systems

Behavioural Issues in Cognitive Systems

An Approach to Intelligent Signal Processing

Matthias Wolff[1] and Rüdiger Hoffmann[2]

[1] Brandenburgische Technische Universität Cottbus,
Lehrstuhl Kommunikationstechnik, 03046 Cottbus, Germany
matthias.wolff@tu-cottbus.de
http://www.tu-cottbus.de/kommunikationstechnik/
[2] Technische Universität Dresden,
Professur Systemtheorie und Sprachtechnologie, 01062 Dresden, Germany
ruediger.hoffmann@tu-dresden.de
http://www.ias.et.tu.dresden.de

Abstract. This paper describes an approach to intelligent signal processing. First we propose a general signal model which applies to speech, music, biological, and technical signals. We formulate this model mathematically using a unification of hidden Markov models and finite state machines. Then we name tasks for intelligent signal processing systems and derive a hierarchical architecture which is capable of solving them. We show the close relationship of our approach to cognitive dynamic systems. Finally we give a number of application examples.

Keywords: intelligent signal processing, hidden Markov automata, hierarchical systems, cognitive systems, acoustic pattern recognition, audio processing.

1 Introduction

In the investigation of multi-modal, verbal and non-verbal human-computer interaction, modeling plays the central role for obtaining engineering solutions. Due to the hierarchical structure of the natural communication behavior, system theory as an engineering discipline has to consider this feature. Therefore we launched a project on building a hierarchical analysis-by-synthesis system more than one decade ago [8]. The progress of this work was presented in the COST 2102 framework [24]. In that time, the concept of cognitive systems was coined [15], and it became clear that our hierarchical approach should be developed towards a hierarchical cognitive dynamic system [23].

In the following, we present the concept of intelligent signal processing as a part of this ongoing research work. Our approach to intelligent signal processing is based on the following paradigm:

> Signals consist of elementary events which are arranged in time and frequency according to a hierarchical pattern.

Some examples shall illustrate that this quite general assumption is reasonable:

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 1–18, 2012.

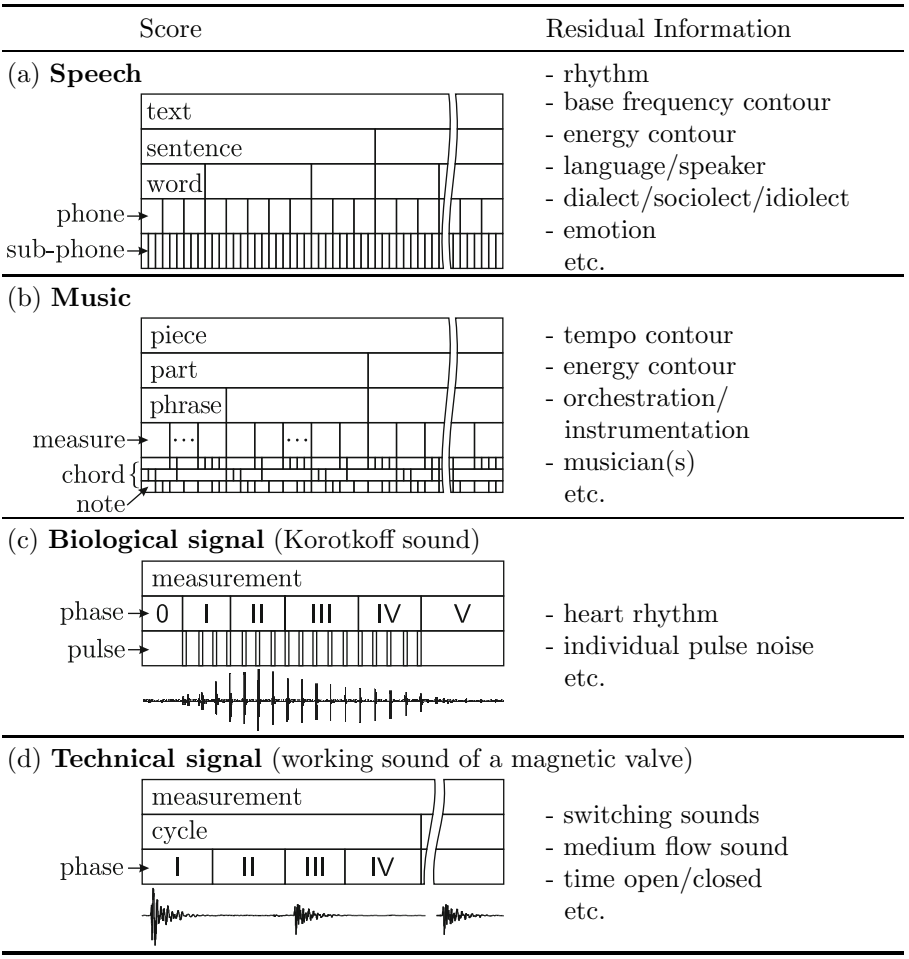


Fig. 1. Examples for the "score" and residual information of signals

Speech. The signal events – from a technical point of view – in speech are sub-phones (usually modeled by HMM states). Sequences of these form phones, syllables, words, sentences, and texts. The symbolic representation of an utterance is a hierarchically organized "score" as shown on the left hand side of Fig. 1a). On each level of the symbol hierarchy the sequencing obeys certain rules which are modeled by lexica, grammars, etc. Of course, the score of the signal only covers a part of the information contained therein (roughly: the linguistic part). There is also para- and non-linguistic information, listed – not claiming to be complete – on the right hand side of Fig. 1a).

Music. In music the signal events are played (or sung) notes. They are arranged in frequency (roughly: chords) and in time forming measures, phrases, parts, and

pieces. The usual symbolic description is a musical score (left hand side of Fig. 1b, the drawing resembles a piano roll). As in speech, the musical score does not cover all the information. Some types of residual information found in musical signals are listed on the right hand side of Fig. 1b).

Biological and Technical Signals. Similar structures can be found in many signals. Figs. 1c and d exemplarily show the "scores" and residual information for a biological signal (Korotkoff sound [30]) and a technical signal (switching sound of a magnetic valve [48]).

We characterize "intelligent signal processing" (ISP) as the capability of discerning score and residual information.[1] Hence, an ISP *system* must be capable of splitting an input signal into these two outputs. This process is usually called *analysis*. An ISP system should also be able to perform the reverse process: *synthesis*. In our paradigm synthesis simply means to compose a signal from its score and sufficient residual information. Fig. 2 shows a black box for an ISP system.

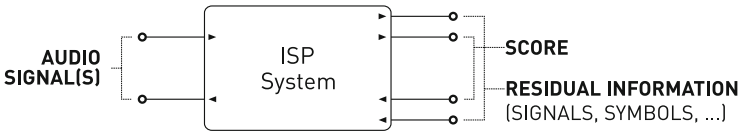


Fig. 2. Intelligent signal processing (ISP) system as a black box

It is obvious that discerning score and residual information requires (application specific) models. These are internal parts of ISP systems and thus not visible in Fig. 2 (cf. Fig. 4 detailing the system architecture). From an information-theoretical view the system is to output only the *information* carried by the signal. The *redundancy* is to be sorted out during analysis and added during synthesis, both basing on the internal models.

In this paper we introduce a mathematical model realizing the paradigm stated above. It unifies hidden Markov models and finite state machines (section 2). Further we propose an architecture for ISP systems (section 3) and describe existing and prospective applications (section 4).

2 Signal Model

2.1 Hidden Markov Automata

We propose *hidden Markov automata* (HMA) as signal models. A hidden Markov automaton is the reformulation of the continuous density hidden Markov model (CD-HMM, [39]) as a finite state transducer (FST, e. g. [32]). The concept was first introduced in [51] and elaborated in [47] and [56]. It is beneficial for the following reasons:

[1] The latin word *intelligere* roughly translates to "to pick-out" or "to discern".

- HMAs provide a single mathematical formalism for all signal processing levels from features to syntax (and even up to semiotics [53,54]).
- Finite state operations like sum, product, closure, composition, etc. are applicable to HMAs.
- The Viterbi and forward algorithms are identical except for the weight semiring [51,47,56].
- The Viterbi (segmental k-means, [28]) and Baum-Welch parameter estimations [1] are identical except for the weight semiring [51,47].
- Likelihood and neglog likelihood computations are identical except for the weight semiring [51,47,56].

A hidden Markov automaton is an octuple

$$\mathcal{H} = \{Z, I, F, \mathcal{O}, Y, S, Q, w\} \tag{1}$$

consisting of

- a finite state alphabet Z,
- a set of initial states $I \subseteq Z$,
- a set of final states $F \subseteq Z$,
- an M-dimensional feature vector space $\mathcal{O} = \mathbb{R}^M$,
- an output alphabet Y,
- a weight semiring $S = (\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$,
- a set of feature mapping functions $Q = \{q_i\}$, $q_i : \mathcal{O} \to \mathbb{K}$, and
- a behavioral function $w : Z \times Q \times Y \times Z \to \mathbb{K}$.

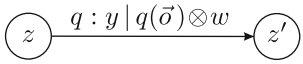
The behavioral function w is usually represented by a list E of transitions e:

$$E = \{e_i\} = \Big\{(z, q, y, z', w)_i\Big\}, \tag{2}$$

each with

- a start state $z \in Z$,
- a feature mapping function $q \in Q$,
- an output symbol $y \in Y$,
- an end state $z' \in Z$, and
- a weight $w \in \mathbb{K}$.

We use the common graphical representation of finite state machines:



(3)

and write $z(e)$, $q(e)$, $y(e)$, $z'(e)$, and $w(e)$ for the elements of the quintuple defining a particular transition e. The function $q(\boldsymbol{o})$ represents the mapping of a feature vector $\boldsymbol{o}$ to the transition.

Hidden Markov automata translate feature vector sequences $\mathbf{o} \in \mathcal{O}^*$ into strings of output symbols $\mathbf{y} \in Y^*$ and assign the weight

$$[\![\mathcal{H}]\!](\mathbf{o}, \mathbf{y}) = \bigoplus_{U \in \mathcal{U}^K(I,\mathbf{y},F)} \left\{ \bigotimes_{e^k \in U} \left[q\left(e^k\right)\left[\boldsymbol{o}^k\right] \otimes w(e^k) \right] \right\} \tag{4}$$

to the translation, where

- $K = |\mathbf{o}|$ is the length of the feature vector sequence,
- $\mathcal{U}^K(I, \mathbf{y}, F)$ denotes the set of consecutive paths through the automaton which have a length of K transitions *and* whose output string is $\mathbf{y}$ (cf. [32]),
- e^k is the k-th transition in such a path,
- $\boldsymbol{o}^k$ is the k-th vector in $\mathbf{o}$,
- $q(e^k)$ is the feature mapping function and $w(e^k)$ is the weight assigned to the transition e^k by the behavioral function w.

Like in HMMs, equation (4) can be computed using the forward algorithm or approximated using the Viterbi algorithm. Both are, however, identical except for the weight semiring S of the HMA [51,47,56]. Table 1 lists the four weight semirings suitable for hidden Markov automata. The feature mapping functions $q(\boldsymbol{o})$ stated with the semirings are appropriately derived from the usual probability density functions (PDF) $p(\boldsymbol{o})$ of hidden Markov models. The transition weights w are derived from the transition probabilities P of HMMs in the same manner.

Table 1. Weight semirings suitable for hidden Markov automata

log.	VITERBI-approx.	Semiring name	$\otimes$	$\oplus$	$q(e)[\boldsymbol{o}^k]$	$w(e)$
no	no	probability	$\cdot$	$+$	$p(\boldsymbol{o}^k \mid q(e))$	$P(e)$
	yes	max/times	$\cdot$	max		
yes	no	logarithmic	$+$	$\oplus_{\ln}$ *)	$-\ln p(\boldsymbol{o}^k \mid q(e))$	$-\ln P(e)$
	yes	tropical	$+$	min		

*) $x \oplus_{\ln} y = -\ln\left(\mathrm{e}^{-x} + \mathrm{e}^{-y}\right)$.

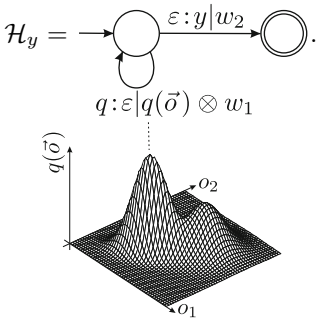
Some differences between HMMs and HMAs are worth mentioning:

- As HMAs translate feature vector sequences into strings of output symbols, the feature mapping functions describe the input rather than the output (as the PDFs do in HMMs). The inversion of an HMA – which translates symbol strings to feature vector sequences – corresponds with the usual HMM interpretation.

– The feature mapping functions are associated with the transitions instead of the states. The concept is known as arc-emission HMM [31].
– HMAs have dedicated initial and final states whereas HMMs have an initial state probability vector and can stop in any state. Even though these concepts are not equivalent, the difference is marginal for the practical application.
– The topology of the automaton graph is not restricted (which is of course not required but still typical for HMMs).

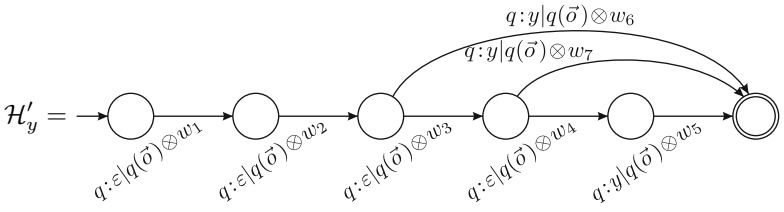
2.2 Event and Score Models

An (idealized) signal event as defined in section 1 manifests as short semi-stationary snippet. Its representation in the feature space is a short sequence of feature vectors with similar statistic properties. Hence we model signal events by elementary hidden Markov automata:



(5)

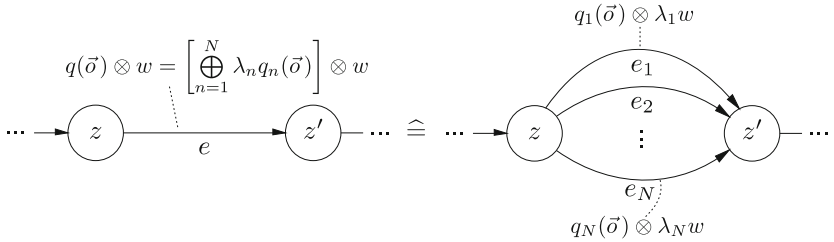
Such an automaton accepts *any* sequence $\mathbf{o}$ of feature vectors, translates it through a single mapping function $q(\boldsymbol{o})$ into exactly one symbol y and assigns a weight $[\![\mathcal{H}_y]\!](\mathbf{o}, y)$ according to equation (4) to the translation. For some technical pattern recognition problems we needed time constraints for the signal events. These can easily be realized. For instance the automaton



(6)

will accept only feature vector sequences of three to five elements.

We derive the mapping functions from Gaussian mixture densities or – which is equivalent in arc emission HMMs – from single Gaussian PDFs:



(7)

A detailed discussion of automata topologies can be found in [56]. The EM parameter estimation for HMA with Gaussian feature mapping functions is described in [47] and [56]. As stated above, the algorithm works for all weight semirings listed in table 1 and unifies the Baum-Welch and the Viterbi (segmental k-means) procedures. [3] and [56] also describe procedures which introduce an automatic topology inference into the EM algorithm.

As explained in section 1, signals can be modeled as an arrangement of elementary events. Hence the simplest score model would be a string. However, such a model is obviously only able to represent one particular score and not "speech" or "music" as a whole. The common solution known from large vocabulary speech processing is to use stochastic sequence models. It was shown by M. Mohri and colleagues that finite state acceptors are suitable for that purpose (e. g. [33,34]). Even though they model only first order Markov chains it is possible to represent also higher order models like n-grams by some simple extensions [35].

We found in our experiments (see section 4) that finite state machines are suitable as score models for other signal types (music, biological and technical signals) as well. In the simplest case we represent the score by a single weighted finite state acceptor $\mathcal{L}$ over the output alphabet Y of the event models. More complex signals like speech or music require a hierarchically organized score model (see section 1). In such cases we describe each level of the hierarchy with a weighted finite state transducer (wFST) $\mathcal{L}_i$ which translates strings from one level to the next higher one. On the highest level $\mathcal{L}_i$ is still an acceptor. The complete score model for all hierarchy levels is obtained through composition:

$$\mathcal{L} = \bigcirc_i \mathcal{L}_i. \tag{8}$$

Fig. 3 shows the proposed model for a particular type of signal.

3 System Design

The signal model suggests a hierarchical processing: first a sub-symbolic level (signal and features) and then at least two symbolic levels (events and score). As indicated above, each of the processing levels needs models to be able to discern

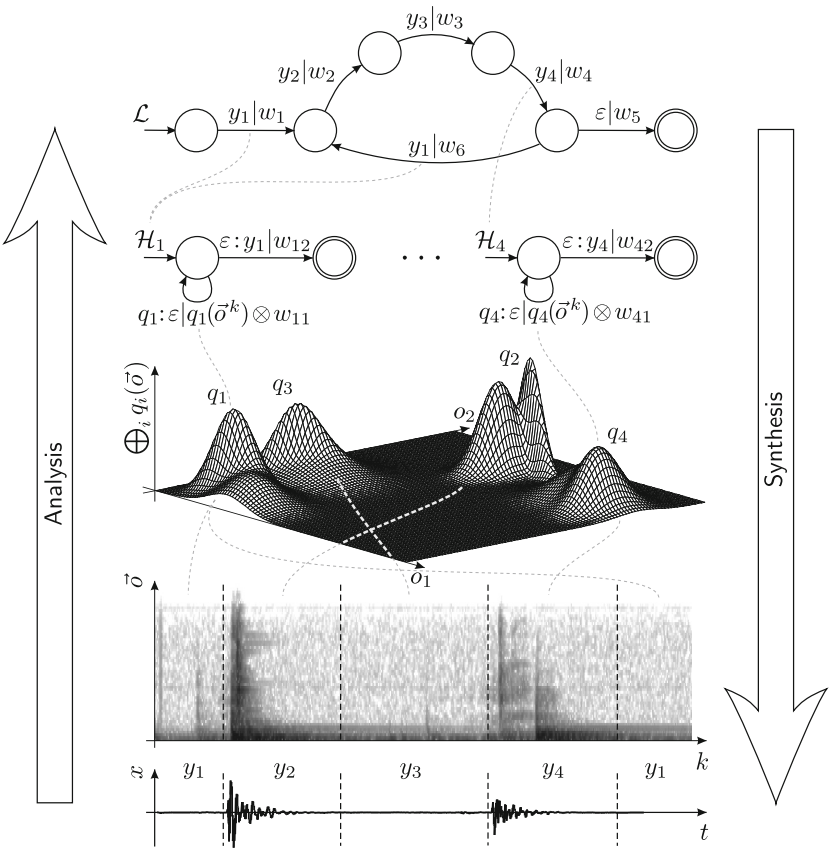


Fig. 3. Example for a signal model (working sound of a magnetic valve, s. Fig. 1d, [47,56]). From bottom to top: signal x with events (y_1="closed", y_2="opening", y_3="open", y_4="closing"), feature vector sequence $\mathbf{o}$ (short-term auto power spectrum), feature mapping functions q_i (two-dimensional projection of the feature space), HMAs $\mathcal{H}_1 \ldots \mathcal{H}_4$ for the events, and score model $\mathcal{L}$ (finite state acceptor).

score, residual information, and redundancy. Obviously, analysis and synthesis use the same models. These simple considerations lead to a design as shown in Fig. 4. The depicted system is a generalization of the "Unified Approach to speech Synthesis and Recognition" (UASR) which we first published in [8] (see also [24,56]). It shall be noted that a very similar "design" was independently found by neuro-biologists in living creatures [14].

The system architecture is symmetrical: each function block on the analysis side (left) has its counterpart on the synthesis side (right). Corresponding analysis and synthesis blocks share a common model. The synthesis algorithm on each level is the inverse of the analysis algorithm and vice versa.

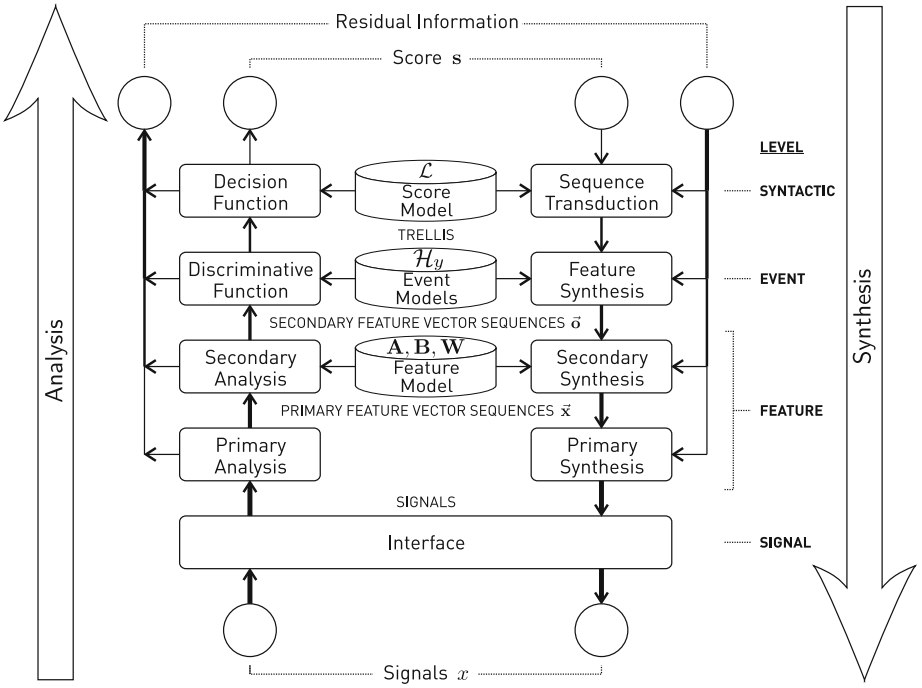


Fig. 4. ISP system design

Interface. This stage performs sensor and actuator (array) processing as well as noise reduction and channel adaptation. Aside from a special speech dereverberation [36] we use standard algorithms.

Primary Analysis and Synthesis. This processing level performs signal analysis and synthesis by standard algorithms (e. g. mel-log spectrum or MFCC for speech, short-term auto power spectrum for technical signals, etc.) It is of course required, that the analysis is reversible (for instance the MLSA synthesis filter [27] approximates the inverse of the mel-log spectrum). Residual information (e. g. fundamental frequency and energy contours in speech) have to be captured by the analysis. The "score" information on this stage is the primary feature vector sequence $\mathbf{x}$:

$$\text{analysis } (x \to \mathbf{x}) : \mathbf{x} = \mathcal{F}\{x\}, \tag{9}$$

$$\text{synthesis } (\mathbf{x} \to x) : x = \mathcal{F}^{-1}\{\mathbf{x}\}, \tag{10}$$

where x denotes the signal, $\mathcal{F}$ denotes the analysis filter, and $\mathcal{F}^{-1}$ its inverse.

Secondary Analysis and Synthesis. The analysis of this stage involves context and dynamic feature computation and the reduction of the feature space dimension. The score-related function can be summarized as follows [47,56]:

$$\text{analysis } (\mathbf{x} \to \mathbf{o}) : \mathbf{o} = \begin{bmatrix}\mathbf{W}_1 \dots \mathbf{W}_n\end{bmatrix} \begin{bmatrix}\mathbf{B}_1(\mathbf{x} - \boldsymbol{x}_0)\mathbf{A}_1 \\ \vdots \\ \mathbf{B}_n(\mathbf{x} - \boldsymbol{x}_0)\mathbf{A}_n\end{bmatrix}, \tag{11}$$

$$\text{synthesis } (\mathbf{o} \to \mathbf{x}) : \mathbf{x} \approx \mathbf{B}_i^{-1}\mathbf{W}_i^{-1}\mathbf{o}\mathbf{A}_1^{-1} + \boldsymbol{x}_0 \quad \text{for } 1 \le i \le n, \tag{12}$$

where

- $\boldsymbol{x}_0$ is an offset vector (normally the mean of the primary features),
- $\mathbf{A}_i$ is a temporal filter matrix,
- $\mathbf{B}_i$ is a spatial filter matrix,[2]
- $\begin{bmatrix}\mathbf{W}_1 \dots \mathbf{W}_n\end{bmatrix}$ is a linear transformation matrix (e. g. PCA or LDA), and
- n is the number of filter operations involved ($n = 3$ is typical for speech processing: 1st - no filtering, 2nd - delta features, 3rd delta-delta features).

We showed in [51,56] that a secondary analysis according to equation (11) is capable of the following functions:

- vector standardization,
- computation of context features,
- computation of difference features,
- temporal and spatial MA-filtering,
- linear feature transformation, and
- reduction of the feature space dimension.

The residual information on this processing stage arises from the dimension reduction of the feature space after the linear transformation. The statistics of the discarded vector components can be kept in order to model the redundancy [5,13]. Thus the residual information may be defined as the difference between the actual values of the components discarded on a particular analysis and the mean vector of those statistics. Both, redundancy and residual, are necessary to exactly restore an individual primary feature vector during synthesis. The statistics alone allow an approximation [5].

The simplified synthesis equation (12) uses only one of the n filtered primary feature vector sequences and does not care for the redundancy and residual information. A solution including the *entire* secondary feature vector $\mathbf{o}$ (but no residual information) is not trivial and requires an optimization. A variant for speech synthesis was presented by Tokuda et al. in [46]. As said before, a solution including only the redundancy was presented in [5].

Event Level. On this level the signal event models $\mathcal{H}_y$ (see section 2.2) are used to translate between feature vector sequences $\mathbf{o}$ and strings $\mathbf{y}$ of events. There is, however, *no* decision made on this stage during analysis. The output of the event level is a trellis which assigns a weight to every possible string of events according to the feature vector sequence (the trellis is of course computed on the fly). The residual information on the event level includes:

[2] The spatial filter matrix was introduced by C. Tschöpe in [47] mainly for performance reasons.

– the duration of the single events,
– the individual characteristics of the signal events. The models $\mathcal{H}_y$ describe a "representative" event only. The individual characteristics are required to re-synthesize the correct signal from the general models. A typical example is speech processing: There are speaker independent acoustic models used for recognition. In order to synthesize a particular speaker from such models, additional speaker adaptation information is required (see e. g. [44]).

Syntactic Level. The syntactic level translates between strings of signal events and a particular score $\mathbf{s}$ through the score model $\mathcal{L}$ (see section 2.2). For analysis this means that a decision for the score is made based on the trellis mentioned above. For synthesis the syntactic stage creates a network of possible event strings for a given score. The residual information on this stage includes variations of the event symbol sequence acceptable for a given score (e. g. pronunciation variants in speech).

The score-related function of the event and syntactic levels can be summarized as follows:

$$\text{analysis } (\mathbf{o} \to \mathbf{s}) : \mathbf{s} = \underset{\mathbf{y} \in Y^*}{\arg\operatorname{ext}} [\![\Big(\bigotimes_{y \in \mathbf{y}} \mathcal{H}_y \Big) \circ \mathcal{L}]\!](\mathbf{o}), \tag{13}$$

$$\text{synthesis } (\mathbf{s} \to \mathbf{o}) : \mathbf{o} \approx \underset{\mathbf{a} \in \mathcal{O}^*}{\arg\operatorname{ext}} [\![\Big(\bigotimes_{y \in \mathbf{s}} \mathcal{H}_y \Big) \circ \mathcal{L}]\!](\mathbf{a}). \tag{14}$$

Like equation (12) the symbolic synthesis according to equation (14) is simplified and does not take care of redundancy and residual information. Thus it only approximates the secondary feature vector sequence.

For the efficient computation of equation (13) we use the Viterbi algorithm or a time-variant A*-search [56] in the HMA

$$\mathcal{R} = \Big(\bigoplus_{y \in Y} \mathcal{H}_y \Big)^* \circ \mathcal{L} \tag{15}$$

which we call a recognition network. For large automata the composition with the score model $\mathcal{L}$ can be done on the fly [2]. Other algorithmic solutions known from speech recognition (like stack decoding) could be applied as well. Until now we have no satisfying general algorithm computing equation (14). As mentioned above, the problem was solved for the special case of speech synthesis by Tokuda et al. An approach to a general solution is made in [41].

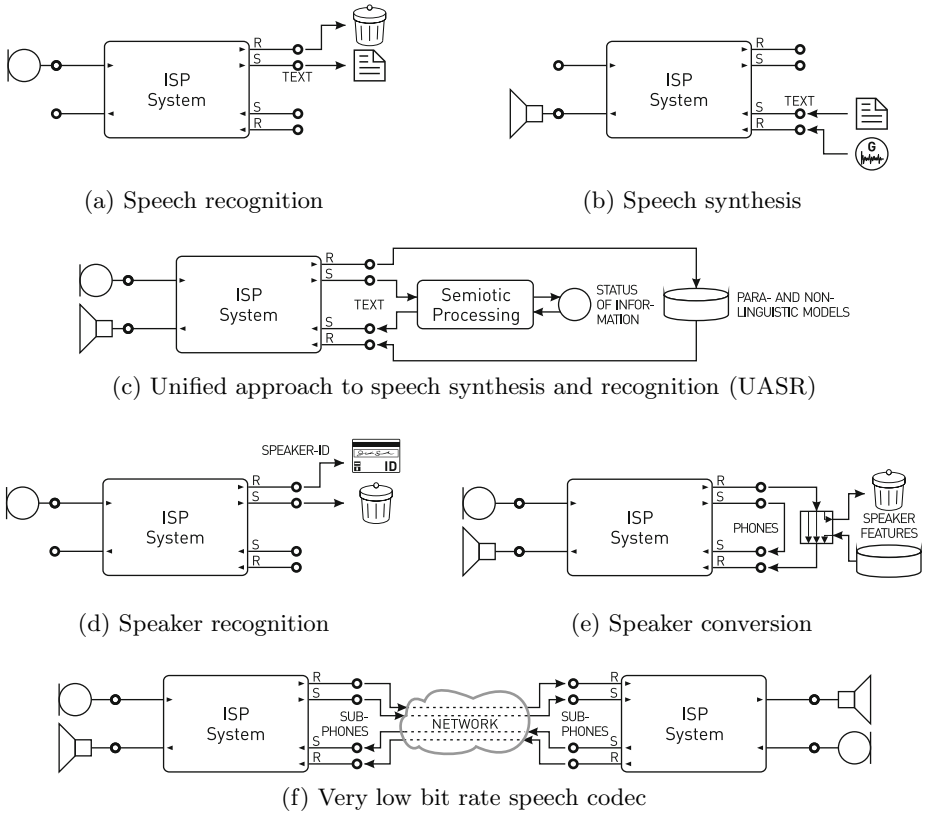
4 Applications

There are a multitude of applications for ISP systems some of which we have realized and experimented with.

Speech. An ISP system can realize a speech recognizer and a speech synthesizer at the same time. The Dresden "Unified Approach to speech Synthesis and Recognition" (UASR, [8,24,56]) is actually the predecessor of this work. Practical applications of the UASR system are described in [4] and [7]. ISP systems can also solve other speech processing tasks like speaker recognition, speaker conversion [44,9], language recognition [52], prosody recognition [29], prosody analysis/synthesis [25], speech coding [43], etc. Fig. 5 shows some "wiring" diagrams for speech applications.

There is of course an abundance of literature on these topics. We would like to point out some ground-breaking work on HMM speech synthesis [60] and very low rate speech coding [45] here.

Music. Our applications of ISP in music signal processing included the identification of individual musical instruments [10,11] and the recognition of genres and styles [17] (Fig. 6c). We also worked on the automatic alignment of lyrics to the music signal [26]. Further applications are musical score recognition (Fig. 6a) and, potentially, a musical synthesizer/sequencer basing on stochastic sound and



(a) Speech recognition

(b) Speech synthesis

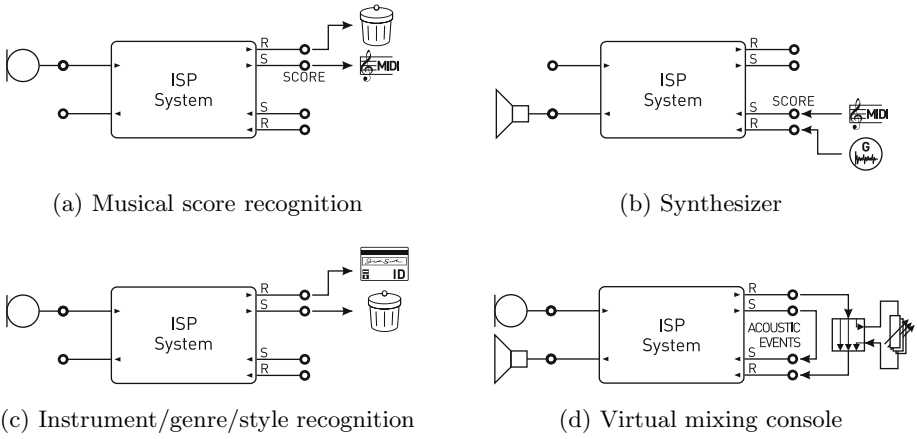
(c) Unified approach to speech synthesis and recognition (UASR)

(d) Speaker recognition

(e) Speaker conversion

(f) Very low bit rate speech codec

Fig. 5. ISP systems in automatic speech processing



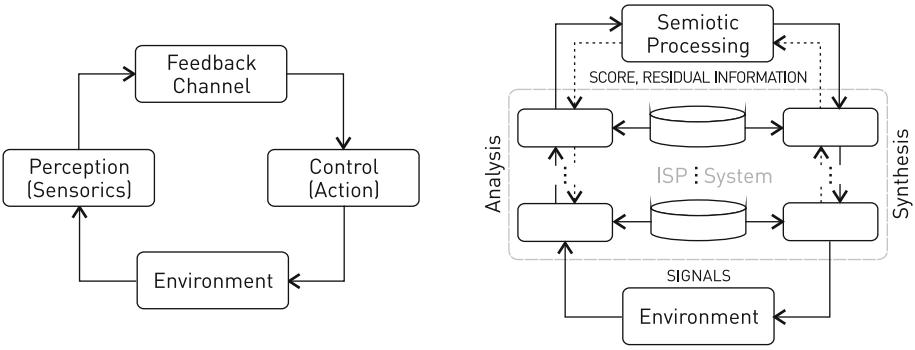
(a) Musical score recognition

(b) Synthesizer

(c) Instrument/genre/style recognition

(d) Virtual mixing console

Fig. 6. ISP systems in music processing



(a) Cognitive dynamic system as defined by S. Haykin in [15,16]

(b) Hierachical cognitive dynamic system incorporating ISP

Fig. 7. Cognitive systems

phrase models (Fig. 6b). The ultimate goal of intelligent music signal processing would be a virtual mixing console (Fig. 6d) which splits arbitrary music signals into their score, timing, expression and instrumentation information, allows the manipulation of all these properties and finally re-synthesizes the manipulated music signal.

Biological and Technical Signals. So far our application of intelligent biological and technical signal processing has been acoustic pattern recognition for various tasks [22,59,40] including: monitoring of train wheels [6], machinery [38,55], and construction elements [48,49,50,51,58], classification of Barkhausen

noise [21], auscultatory blood pressure measurement [57], classification of chewing sounds [37], and quality assessment of tissue papers [59]. Some special solutions have been patented [12,18,19,20].

The main goal of our future research on this field will be the construction of cognitive systems [15] for controlling machinery, power grids, etc. The "synthesis" part in such systems generates control signals which influence the system's environment in a target-oriented way. We are also investigating the use of simulated or synthesized sensor signals for model training and verification.

5 Conclusion

We proposed a conceptional and mathematical model of signals and presented a technical specification of a system for intelligent signal processing. The system has been realized in software and – for speech – also in hardware (DSP & FPGA, [4]). We presented practical applications for speech, music, and acoustic pattern recognition. The Dresden "Unified Approach to Speech Synthesis and Recognition" [24] is an instance of an intelligent signal processing system.

Our current research focuses on the extension of our approach into a hierarchical cognitive dynamic system (Fig. 7 illustrates the relationship). ISP constitutes the "perception" and "action" blocks of a cognitive dynamic system. We consider the following properties of our ISP system as crucial for actual cognitive capabilities:

- hierarchical processing (cf. [14], also stressed by S. Haykin in [16]),
- shared models on all hierarchy levels (also cf. [14]), and
- bidirectional processing (dotted arrows in Fig. 7b; this feature not elaborated in this paper, see [42] in this book).

Our approach to complete the IAP system for speech processing by a semiotic processing ("feedback channel" in Fig. 7a) is described in this book [54].

Acknowledgments. This work was partially founded by

- the Deutsche Forschungsgemeinschaft (DFG) under grants Ho 1674/3, Ho 1674/7, and Ho1684/8,
- the German Federal Ministry of Education and Research (BMBF) under grants 01 IV 701 K2, 01 IV 102 L8, 01 NM 136 C, 03 i 4745 A, and 13 N 9793, and
- the Arbeitsgemeinschaft industrieller Forschungsvereinigungen "Otto von Guericke" (AiF) under grants KF 0413401 WD6 and KF 2282501 WD9.

References

1. Bilmes, J.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and hidden Markov models. Tech. rep., International Computer Science Institute (1998)

2. Caseiro, D., Trancoso, I.: A specialized on-the-fly algorithm for lexicon and language model composition. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1281–1291 (2006)
3. Duckhorn, F.: Optimierung von Hidden-Markov-Modellen für die Sprach- und Signalerkennung. Diplomarbeit, Technische Universität Dresden, Institut für Akustik und Sprachkommunikation (2007)
4. Duckhorn, F., Wolff, M., Strecha, G., Hoffmann, R.: An application example for unified speech synthesis and recognition using Hidden Markov Models. In: One Day Meeting on Unified Models for Speech Recognition and Synthesis, Birmingham, U.K. (March 2009)
5. Eichner, M.: Spracherkennung und Sprachsynthese mit gemeinsamen Datenbasen - Akustische Analyse und Modellierung. Dissertationsschrift, Technische Universität Dresden, Institut für Akustik und Sprachkommunikation, Studientexte zur Sprachkommunikation vol. 43, w.e.b. Universitätsverlag, Dresden (2006) ISBN 978-3-940046-10-9
6. Eichner, M.: Signalverarbeitung für ein rotationsbezogenes Messsystem. Forschungsbericht, Technische Universität Dresden, Institut für Akustik und Sprachkommunikation (April 2007)
7. Eichner, M., Göcks, M., Hoffmann, R., Kühne, M., Wolff, M.: Speech-enabled services in a web-based e-learning environment. Advanced Technology for Learning 1(2), 91–98 (2004)
8. Eichner, M., Wolff, M., Hoffmann, R.: A unified approach for speech synthesis and speech recognition using Stochastic Markov Graphs. In: Proceedings of the Internation Conference on Spoken Language Processing, ICSLP 2000, Beijing, PR China, vol. 1, pp. 701–704 (October 2000)
9. Eichner, M., Wolff, M., Hoffmann, R.: Voice characteristics conversion for TTS using reverse VTLN. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Canada, vol. 1, pp. 17–20 (May 2004)
10. Eichner, M., Wolff, M., Hoffmann, R.: Instrument classification using Hidden Markov Models. In: International Conference on Music Information Retrieval, ISMIR 2006, Victoria, BC, Canada, pp. 349–350 (October 2006)
11. Eichner, M., Wolff, M., Hoffmann, R.: An HMM based investigation of differences between musical instruments of the same type. In: Proceedings of the International Congress on Acoustics, ICA 2007, Madrid, Spain, 5 pages on CD-ROM Proceedings (September 2007)
12. Eichner, M., Wolff, M., Hoffmann, R., Kordon, U., Ziegenhals, G.: Verfahren und Vorrichtung zur Klassifikation und Beurteilung von Musikinstrumenten. Deutsches Patent 102006014507 (December 2008)
13. Eichner, M., Wolff, M., Ohnewald, S., Hoffmann, R.: Speech synthesis using stochastic Markov graphs. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2001, Salt Lake City, UT, USA, pp. 829–832 (May 2001)
14. Fuster, J.M.: Cortex and Mind: Unifying Cognition. Oxford University Press, New York (2005) 978-0-19-530084-0
15. Haykin, S.: Cognitive dynamic systems. Proceedings of the IEEE 94(11), 1910–1911 (2006)
16. Haykin, S.: Foundations of cognitive dynamic systems. IEEE Lecture, Queens University (January 29, 2009), http://soma.mcmaster.ca/papers/Slides_Haykin_Queens.pdf

17. Hübler, S.: Suchraumoptimierung zur Identifizierung ähnlicher Musikstücke. Diplomarbeit, Technische Universität Dresden, Institut für Akustik und Sprachkommunikation (2008)
18. Hentschel, D., Tschöpe, C., Hoffmann, R., Eichner, M., Wolff, M.: Verfahren zur Beurteilung einer Güteklasse eines zu prüfenden Objekts. Deutsches Patent 10 2004 023 824 (July 2006)
19. Hentschel, D., Tschöpe, C., Hoffmann, R., Eichner, M., Wolff, M.: Verfahren zur Beurteilung einer Güteklasse eines zu prüfenden Objekts. Europäisches Patent EP 1 733 223 (January 2008)
20. Hentschel, D., Tschöpe, C., Hoffmann, R., Eichner, M., Wolff, M.: Verfahren zur Beurteilung einer Güteklasse eines zu prüfenden Objekts. Österreichisches Patent AT 384261 (February 2008)
21. Erkennungsexperimente mit Barkhausen-Rauschen. In: Hoffmann, R. (ed.) Jahresbericht 1999, p. 34. Technische Universität Dresden, Institut für Akustik und Sprachkommunikation (December 1999)
22. Hoffmann, R.: Recognition of non-speech acoustic signals. In: Kacic, Z. (ed.) Proceedings of the International Workshop on Advances in Speech Technology Advances, AST 2006, p. 107. University of Maribor, Maribor (2006)
23. Hoffmann, R.: Denken in Systemen. In: Gerlach, G., Hoffmann, R. (eds.) Neue Entwicklungen in der Elektroakustik und elektromechanischen Messtechnik, Dresdner Beiträge zur Sensorik, vol. 40, pp. 13–24. TUD Press, Dresden (2009)
24. Hoffmann, R., Eichner, M., Wolff, M.: Analysis of Verbal and Nonverbal Acoustic Signals with the Dresden UASR System. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) Verbal and Nonverbal Commun. Behaviours. LNCS (LNAI), vol. 4775, pp. 200–218. Springer, Heidelberg (2007)
25. Hussein, H., Strecha, G., Hoffmann, R.: Resynthesis of prosodic information using the cepstrum vocoder. In: Proceedings of the 5th International Conference Speech Prosody. Chicago, IL, March 11-14, 4 pages (2010)
26. Hutschenreuther, T.: Automatische Anordnung von Gesangstexten zu Musik mit Hilfe von Methoden aus der Spracherkennung. Diplomarbeit, Technische Universität Dresden, Institut für Akustik und Sprachkommunikation (2009)
27. Imai, S., Sumita, K., Furuichi, C.: Mel log spectrum approximation (MLSA) filter for speech synthesis. In: Electronics and Communications in Japan (Part I: Communications), vol. 66, pp. 10–18 (1983)
28. Juang, H.H., Rabiner, L.R.: The segmental K-means algorithm for estimating parameters of Hidden Markov Models. IEEE Transactions on Acoustics, Speech, Signal Processing 38(9), 1639–1641 (1990)
29. Kühne, M., Wolff, M., Eichner, M., Hoffmann, R.: Voice activation using prosodic features. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2004, pp. 3001–3004 (October 2004)
30. Korotkoff, N.C.: On the subject of methods of determining blood pressure. Bull. Imperial. Mil. Med. Acad. 11, 365–367 (1905)
31. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (2001)
32. Mohri, M.: Weighted automata algorithms. In: Droste, M., Kuich, W., Vogler, H. (eds.) Handbook of Weighted Automata. Monographs in Theoretical Computer Science. An EATCS Series, pp. 213–254. Springer, Heidelberg (2009) ISBN 978-3-642-01491-8
33. Mohri, M., Pereira, F., Riley, M.: Speech recognition with weighted finite-state transducers. In: Handbook on Speech Processing and Speech Communication, Part E: Speech Recognition. Springer (2008)

34. Mohri, M., Riley, M.: Weighted finite-state transducers in speech recognition (tutorial). In: Proceedings of the International Conference on Spoken Language Processing (2002)
35. Mohri, M., Riley, M., Hindle, D., Ljolje, A., Pereira, F.: Full expansion of context-dependent networks in large vocabulary speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1998, vol. 2, pp. 665–668 (May 1998)
36. Petrick, R., Lohde, K., Wolff, M., Hoffmann, R.: The harming part of room acoustics in automatic speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2007, Antwerp, Belgium, pp. 1094–1097 (August 2007)
37. Päßler, S., Wolff, M., Fischer, W.J.: Chewing sound classification using a grammar based classification algorithm. In: Proceedings of Forum Acusticum 2011 (2011) ISBN 978-84-694-1520-7
38. Pusch, T., Cherif, C., Farooq, A., Wittenberg, S., Hoffmann, R., Tschöpe, C.: Early fault detection at textile machines with the help of structure-borne sound analysis. Melliand English 11-12, E144–E145 (2008)
39. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
40. Richter, T.: Erkennung von Biosignalen. Diplomarbeit, Technische Universität Dresden, Institut für Akustik und Sprachkommunikation (2001)
41. Römer, R.: Beschreibung von Analyse-Synthese-Systemen unter Verwendung von kaskadierten bidirektionalen HMMs. In: Kröger, B.J., Birkholz, P. (eds.) Elektronische Sprachsignalverarbeitung 2011, Tagungsband der 22. Konferenz. Studientexte zur Sprachkommunikation, vol. 61, pp. 67–74. TUD Press (2011) ISBN 978-3-942710-37-4
42. Römer, R.: A Cortical Approach Based on Cascaded Bidirectional Hidden Markov Models. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) Cognitive Behavioural Systems. LNCS, vol. 7403, pp. 266–272. Springer, Heidelberg (2012)
43. Strecha, G., Wolff, M.: Speech synthesis using hmm based diphone inventory encoding for low-resource devices. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), pp. 5380–5383 (2011)
44. Strecha, G., Wolff, M., Duckhorn, F., Wittenberg, S., Tschöpe, C.: The HMM synthesis algorithm of an embedded unified speech recognizer and synthesizer. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, Brighton, U.K., pp. 1763–1766 (September 2009)
45. Tokuda, K., Masuko, T., Hiroi, J., Kobayashi, T., Kitamura, T.: A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 609–612 (1998)
46. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 1315–1318 (2000)
47. Tschöpe, C.: Klassifikation technischer Signale, Studientexte zur Sprachkommunikation, vol. 60. TUD Press (2012)

48. Tschöpe, C., Hentschel, D., Wolff, M., Eichner, M., Hoffmann, R.: Classification of non-speech acoustic signals using structure models. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, vol. 5, pp. V653–V656 (May 2004)
49. Tschöpe, C., Hirschfeld, D., Hoffmann, R.: Klassifikation technischer Signale für die Geräuschdiagnose von Maschinen und Bauteilen. In: Tschöke, H., Henze, W. (eds.) Motor- und Aggregateakustik II, pp. 45–53. Expert Verlag, Renningen (2005)
50. Tschöpe, C., Wolff, M.: Automatic decision making in SHM using Hidden Markov Models. In: Database and Expert Systems Applications, DEXA 2007, pp. 307–311 (September 2007)
51. Tschöpe, C., Wolff, M.: Statistical classifiers for structural health monitoring. IEEE Sensors Journal 9(11), 1567–1676 (2009)
52. Werner, S., Wolff, M., Eichner, M., Hoffmann, R., Estelmann, J.: Language identification using meta-classification of multiple experts. In: Processings of the International Conference on Speech and Computer, SPECOM 2005, Patras, Greece, pp. 519–522 (October 2005)
53. Wirsching, G., Huber, M., Kölbl, C.: The confidence-probability semiring. Tech. Rep. 2010-4, Institut für Informatik der Universität Augsburg (2010)
54. Wirsching, G., Huber, M., Kölbl, C., Lorenz, R., Römer, R.: Semantic Dialogue Modeling. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) Cognitive Behavioural Systems. LNCS, vol. 7403, pp. 104–113. Springer, Heidelberg (2012)
55. Wittenberg, S., Wolff, M., Hoffmann, R.: Feasibility of statistical classifiers for monitoring rollers. In: Proceedings of the International Conference on Signals and Electronic Systems, ICSES 2008, Krakow, Poland, pp. 463–466 (September 2008)
56. Wolff, M.: Akustische Musterkennung, Studientexte zur Sprachkommunikation, vol. 57. TUD Press (2011) ISBN 978-3-942710-14-5
57. Wolff, M., Kordon, U., Hussein, H., Eichner, M., Hoffmann, R., Tschöpe, C.: Auscultatory blood pressure measurement using HMMs. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, Honolulu, HI, USA, vol. 1, pp. 405–408 (April 2007)
58. Wolff, M., Schubert, R., Hoffmann, R., Tschöpe, C., Schulze, E., Neunübel, H.: Experiments in acoustic structural health monitoring of airplane parts. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, Las Vegas, NV, USA, pp. 2037–2040 (April 2008)
59. Wolff, M., Tschópe, C.: Pattern recognition for sensor signals. In: Proceedings of the IEEE Sensors Conference 2009, Christchurch, New Zealand, pp. 665–668 (October 2009)
60. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication 51(11), 1039–1154 (2009)

The Analysis of Eye Movements in the Context of Cognitive Technical Systems: Three Critical Issues

Sebastian Pannasch[1,2], Jens R. Helmert[2], Romy Müller[2], and Boris M. Velichkovsky[2,3]

[1] Brain Research Unit, Aalto University School of Science, Espoo, Finland
[2] Applied Cognitive Research/Psychology III, Technische Universitaet Dresden, Germany
[3] Department of Cognitive Studies, Kurchatov Institute, Moscow, Russian Federation
{pannasch,helmert,mueller,velich}@applied-cognition.org

Abstract. Understanding mechanisms of attention is important in the context of research and application. Eye tracking is a promising method to approach this question, especially for the development of future cognitive technical systems. Based on three examples, we discuss aspects of eye gaze behaviour which are relevant for research and application. First, we demonstrate the omnipresent influence of sudden auditory and visual events on the duration of fixations. Second, we show that the correspondence between gaze direction and attention allocation is determined by characteristics of the task. Third, we explore how eye movements can be used for information transmission in remote collaboration by comparing it with verbal interaction and the mouse cursor. Analysing eye tracking in the context of future applications reveals a great potential but requires solid knowledge of the various facets of gaze behavior.

Keywords: eye movements, attention, fixation duration, remote collaboration.

1 Introduction

Action and interaction with objects and other persons in the environment requires attention. The definition, understanding, and measurement of attention is one of the central research topics in psychology and cognitive science [1]. This interest is motivated not only by fundamental research questions but also by the increasing complexity of our (technical) environments. Currently, it becomes more and more challenging for users of technical devices to monitor and organize their interaction with them, as these requirements strongly increase mental load. Future developments, therefore, should build on solid knowledge about perception, attention and information processing to directly incorporate these processes into the design of attention- and intention-sensitive interfaces.

To approach this problem, several behavioural and psychophysiological measurement techniques have been employed in the past. Within the methodological arsenal, eye tracking and the analysis of human eye movements are most appropriate to investigate attention and provide attention-based support. This argument is founded on two main advantages of eye tracking. First, it is assumed that the direction of the

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 19–34, 2012.

eyes corresponds to the allocation of visual attention, thus measuring gaze behaviour can provide insights about mental processing [2]. Second, video-based eye tracking is a non-invasive, general tool, which can be applied to almost all everyday life situations [3].

In humans, as in all higher primates, vision is the dominant sensory modality and the important role of eye movements for visual processing has been repeatedly emphasized [e.g. 4]. During visual perception, information is sampled from the environment via 'active vision' [5]. Saccades—fast ballistic movements—redirect the foveal region of the eyes from one fixation point to another. During saccades, the intake and processing of visual information is largely suppressed and therefore limited to the periods of fixations [6]. This interplay of fixations and saccades is essential, as highest visual acuity is limited to the small foveal region; outside this high-resolution area, vision becomes blurred and the perception of colour is reduced. Eye movement behaviour in many everyday situations, such as reading text or inspecting images, can be described as an alternation between fixations and saccades.

Investigating human eye movement behaviour generates a quantity of rich data and therefore allows for an analysis of various parameters [for reviews 7, 8]. Traditionally, these analyses have largely relied on *when*, *where* and *how* information is gathered from the visual environment. Here, we will focus on the first two aspects and additionally consider a specific feature of gaze behaviour in *social* interaction. First, considering the *when* aspect is of importance with regard to the level of information processing. Particularly, we will examine the duration of fixations in free visual exploration. Second, regarding the *where* characteristic of gaze behaviour, it is usually assumed that the direction of the eyes allow for accurate estimations of the ongoing focus of interest and processing at any given time. Third, when communicating with other people, gaze behaviour has a particular function in *social* interaction; here we will investigate its contribution when direct communication is impaired in situations of remote collaboration. The selected characteristics provide representative examples about the functional importance of eye movements when trying to understand mechanisms of attention and information processing.

In our opinion, eye tracking will play an important role in the development of attentive interfaces. First versions of technical systems based on the analysis of eye gaze behaviour have already emerged [e.g. 9]. However, it should be mentioned here that this perspective is not new, and in each decade since the 1950s, the discussion about the use of eye tracking for solving new problems has persistently returned [10]. While confident of the potential of eye tracking, we are nonetheless well aware of the potential risks for failure when eye movements are assumed to serve as a simple attention pointer. In fact, successfully implementing attention-sensitive devices requires a deep understanding of the underlying mechanisms of gaze control as well as a careful interpretation of the resulting behaviour.

In the following sections we will present recent results from three different domains of eye movement research and thereby highlight potentials and pitfalls in understanding the complex control mechanisms of eye movements and discuss their significance for the development of cognitive technical systems.

2 Sensitivity of Fixations to Distraction

Visual fixations represent the time intervals dedicated to visual information uptake and processing. Their durations are often considered as reflecting the entropy of the fixated information: longer fixation durations are associated with the processing of demanding information and higher task complexity [11]. It is generally assumed that the employment of more cognitive efforts is expressed in longer fixations, for instance when eye movements are analysed in the context of reading [12] or scene perception [13]. However, this hypothesis raises the question if and to what extent also other factors can modulate the duration of fixations, as this would make it difficult to attribute these temporal variations to the ongoing information processing. In fact, it has been shown that sudden changes in the environment lead to a robust prolongation of the fixation duration [e.g. 14].

Understanding the underlying mechanisms of this change-related prolongation is an interesting research endeavour in itself, but it also turns out to be of particular importance when analysing eye movements in applied contexts. For instance, it has been demonstrated that the change of a traffic light (i.e. from green to red) results in a pronounced prolongation of the respective fixation [15]; based on such a feature, one could think of fixation-based hazard recognition.

Therefore, it is necessary to understand what processes take place within a single fixation and how they contribute to its duration. In the present experiment, we examined if periods of higher or lower sensitivity to distraction within a fixation can be identified. Recently, it has been reported that fixations can be influenced by the appearance of visual, acoustic and haptic events [16]. To further investigate this phenomenon, we presented visual and auditory distractors.

2.1 Methods

Subjects. Seventeen students (10 females) of the Technische Universität Dresden with a mean age of 23.4 years (range 20-30 years) took part in this experiment. All subjects reported normal or corrected-to-normal vision, normal hearing and received course credit for participation. The study was conducted in conformity with the declaration of Helsinki.

Apparatus. Participants were seated in a dimly illuminated, sound-attenuated room. Eye movements were sampled monocularly at 250 Hz using the SR EyeLink I infrared eye tracking system with on-line detection of saccades and fixations and a spatial accuracy of better than 0.5°. Stimuli were shown using a CRT display (19-inch Samtron 98 PDF) at 800 by 600 pixels at a refresh rate of 100 Hz. Viewed from a distance of 80 cm, the screen subtended a visual angle of 27.1° horizontally and 20.5° vertically.

Stimuli. Ten digitized pieces of fine art by European seventeenth to nineteenth-century painters served as stimulus material. Visual and auditory distractors were presented to systematically investigate influences of gaze-contingent distractions. Visual distractors were implemented as colour inversion of an image segment with a

size of 50 by 50 pixels, always appearing 50 pixels to the left of the ongoing fixation. Auditory distractors consisted of a 900 Hz sinusoidal tone, presented at a sound pressure level of 70 dB via PC-loudspeakers on both sides of the screen.

Procedure. Subjects were informed that the purpose of the study was to investigate eye movement patterns in art perception and were asked to study the images in order to be prepared to answer subsequent questions regarding the image content. They were aware of the presentation of distractors but instructed to ignore them. The experiment was run in two consecutive blocks of varying distractor modality (visual or auditory), each containing five pictures. The order of blocks was counterbalanced across subjects. A 9-point calibration and validation was performed before the start of each block. Before each trial, a drift correction was performed. Distractor presentation always began after an initial period of 20 s of scene inspection in order to allow subjects firstly to explore each image without disturbance. Once all 21 distractors (see below) were shown, the image was replaced by five questions which had to be answered by clicking 'yes' or 'no' on-screen buttons using the mouse. The total duration of the experiment was about 40 min.

Distractors were presented at every fifth fixation during a trial. This presentation interval was selected according to previous work [16] and warranted enough unaffected fixations in between, serving as baseline. Distractors were triggered by the fixation onset with a latency of 50, 100, 150, 200, 250, 300 or 350 ms and presented with a duration of 75 ms. For each onset delay, three distractors were shown in a randomized order, resulting in a total of 21 distractors per image. If a fixation was terminated before reaching the onset latency, the program waited for the next suitable fixation. The image presentation lasted until all 21 distractors were presented (65 seconds on average).

2.2 Results

Fixations around eyeblinks and outside the presentation screen were removed. Further processing included only distracted fixations and the two adjacent non-distracted fixations. The non-distracted fixations served as baseline. To assure comparability of the baseline and the distractor condition, fixations of shorter duration than the respective distractor latency (see above) were excluded, resulting in a total of 25686 (82%) valid fixations. *Eta*-squared values are reported as estimates of the effect size [17].

To investigate the effects of the visual and auditory distractors, fixation durations of the baseline and the distractor condition were compared. Medians of fixation duration were applied to a 2 (modality: visual, auditory) × 2 (fixation type: distracted, baseline) repeated measures analysis of variance (ANOVA) and revealed significant main effects for fixation type, $F(1,16) = 223.96$, $p < .001$, $\eta^2 = .68$, but not for modality, $F < 1$. Furthermore, we found a significant interaction for modality × fixation type, $F(1,16) = 5.95$, $p = .027$, $\eta^2 = .004$. Regarding the main effect of fixation type, fixation durations were longer when affected by a distractor presentation (*M*s: 315 vs. 250 ms). The interaction was based on the slightly stronger influence of visual distractors

(*M*s: 318 vs. 311 ms), while the average fixation duration in both baseline conditions was similar (*M*s: 248 vs. 251 ms).

In order to examine if the appearance of a distraction at various latencies within a fixation induces differential effects, we calculated the differences between distracted and baseline fixations, for each modality and latency. The obtained difference values were applied to a 2 (modality: visual, auditory) × 7 (latency: 50, 100, 150, 200, 250, 300, 350) repeated measures ANOVA and revealed a significant main effect for modality, $F(1,16) = 20.05$, $p < .001$, $\eta^2 = .084$, but not for latency, $F(6,96) = 1.69$, $p = .132$. No interaction effect was found, $F < 1$. The obtained main effect for modality is based on a stronger general influence of visual distraction, evidenced in larger difference values between baseline and distractor fixations (*M*s: 46 vs. 13 ms).

2.3 Discussion

In accordance with previous findings [16], we observed event-related prolongations of fixations for visual as well as for auditory distraction. This result is important for the interpretation of fixation durations in applied contexts: Something as ordinary as a ringing phone might be responsible for a prolonged fixation. Thus, the fixation duration represents a highly sensitive parameter, reflecting internal processing mechanisms as well as reacting to external events. Incorporating the fixation duration in attention-sensitive interfaces therefore requires considering this interaction. The analysis of the difference values provides clear evidence that the appearance of a distracting event at any time within a fixation evokes a similar prolongation effect.

Furthermore, the influence of visual distractors was stronger than that of auditory distractors, which corresponds to earlier reports [16, 18]. However, based on the current findings, it remains to be determined whether this difference results from different processing mechanisms or to a lack of comparability between the two types of distractors. Although both distracting events were shown within the same experimental paradigm, we cannot be sure that a colour inversion of 50 x 50 pixels represents a comparable event to a 900 Hz sinusoidal tone of 70 dB.

3 Focus of Attention

In general, it is assumed that visual attention is allocated to the position of the current fixation. Reportability of fixational content is often considered a measure of attention allocation. While in most cases a perfect fit of gaze position and attentional direction can be found, there is evidence that subjects report contents ahead [19] as well as behind [20] the position of the current fixation. It therefore is of interest to understand if the relationship between fixation position and attention allocation changes according to particular requirements (for instance with respect to the task at hand) or if the above mentioned controversial results are rather based on differences in the paradigms.

Support for the attention-ahead-of-fixation assumption is mainly found in laboratory settings using so-called ‘fixate-and-jump’ paradigms. In such settings, saccades

are programmed due to arbitrary commands, as criticized by Fischer [21]. In complex, everyday tasks under rather natural settings, it has been shown that subjects report the content of the current or previous fixation. Notwithstanding, to employ eye movement analysis in applied domains, these temporal characteristics are essential and require a deeper understanding. To contribute to this discussion and to allow for a precise investigation of attention allocation, we analysed the subjective focus of visual attention in a continuous paradigm using different task instructions. The current experiment is based on hierarchical approaches of attention, assuming that attention operates on different levels [e.g. 1, 11, 22].

3.1 Method

Subjects. Eighteen students (9 females) of the Technische Universität Dresden with a mean age of 23.4 years (range 20-31 years) took part in this experiment. All subjects reported normal or corrected-to-normal vision, normal hearing and received course credit for participation in the study conducted in conformity with the declaration of Helsinki.

Apparatus. The same apparatus as in the first experiment was used.

Stimuli. A total of 121 black and white pictograms served as stimuli and were shown with a size of 2° x 2° of visual angle. Within each trial, six pictograms were presented in a circular array with a diameter of 13.3° of visual angle (Figure 1C).

Procedure. All subjects completed three blocks of 60 trials. Within one block all trials were of the same task condition. Three different tasks were employed. In the *position* condition (Figure 1A), the task was to click on the empty position where subjects felt they were looking at the time of the trial end. In the *content* condition, subjects had to choose the pictogram they thought they had inspected during the beep. Therefore, the test screen contained three pictograms: the actual fixated one, the previously fixated one and the next one in the array (Figure 1B). In the condition *position and content*, the trial screen remained unchanged.

A 9-point calibration and validation was performed before the start of the experiment. Each trial started with a drift correction at one of the six locations where the pictograms were shown. Among the trials, the drift correction location was

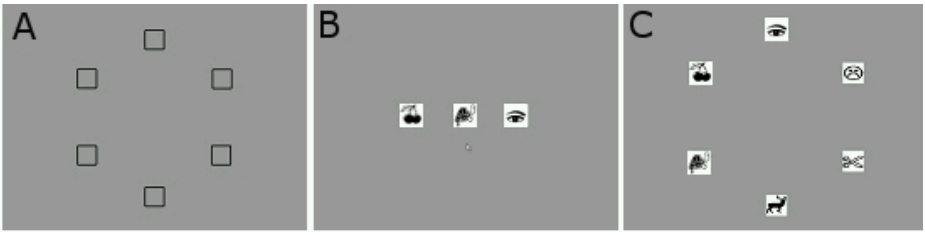


Fig. 1. Initial arrangement of pictograms during the trials (C). Test screens for the different tasks: Position (A), Content (B), and Position and Content (C).

randomly selected but counterbalanced across the block. After the onset of the pictograms, subjects had to scan the display in a clockwise manner, starting from the drift correction position. As soon as a predefined pictogram was fixated, a countdown started (400-600 ms, steps of 50 ms). Once the end of the countdown was reached, a beep was provided to signal the end of presentation. This countdown procedure and time ensured that the trial end in the majority of cases appeared around the start and end of a fixation. Subsequently, the test screens were presented. After the judgments were made, a new trial was initiated. The presentation order of blocks was counterbalanced across subjects; at the end of each block, subjects had a break of five minutes. An experimental session lasted approximately 40 minutes.

3.2 Results

Before the statistical analysis, some trials were rejected due to invalid recording. Furthermore, trials were excluded in which the last fixation position did not correspond to the position of a pictogram. Subsequent to this preprocessing, 2380 valid trials remained (74% of all trials). In these valid trials, participants had chosen the pictogram they felt they last fixated: the pictogram previous to the one the eye fixated during the signal tone (*previous*); the actual pictogram the eye fixated during the signal tone (*actual*), or the next pictogram (*next*). As another factor in the analyses, the viewing times of the last pictogram were considered. They were divided into three categories based on tertiles, with the same number of cases in each category. This resulted in three viewing time conditions; the respective median values and ranges are shown in Table 1.

Table 1. Tertile-based categories of viewing times (in ms)

Viewing Time	Minimum	Median	Maximum	N
Short	1	73	139	795
Medium	140	243	400	824
Long	401	498	601	761

The viewing time categories (short, middle, long) served as independent variables for further statistical testing. The dependent variables were probabilities of *previous* and *next* responses. Probabilities of choosing the *actual* position/pictogram were not analysed, as these cases indicated an overlap of eye fixation and perceived focus of visual attention, hence providing no diagnostic information. Two 3 (viewing time) × 3 (task) repeated measures ANOVAs were conducted. For *previous*, significant main effects for task, $F(2, 34) = 3.90$, $p = .030$, $\eta^2 = .02$, and viewing time, $F(2, 34) = 25.23$, $p < .001$, $\eta^2 = .34$, as well as a significant interaction, $F(4, 68) = 6.90$, $p < .001$, $\eta^2 = .05$, were obtained. Concerning task, results show that the highest probability for the *previous* choice was in the *content* condition (11.1%), followed by *position & content* (9.0%) and *position* (5.2%). The factor viewing time clearly shows a decrease

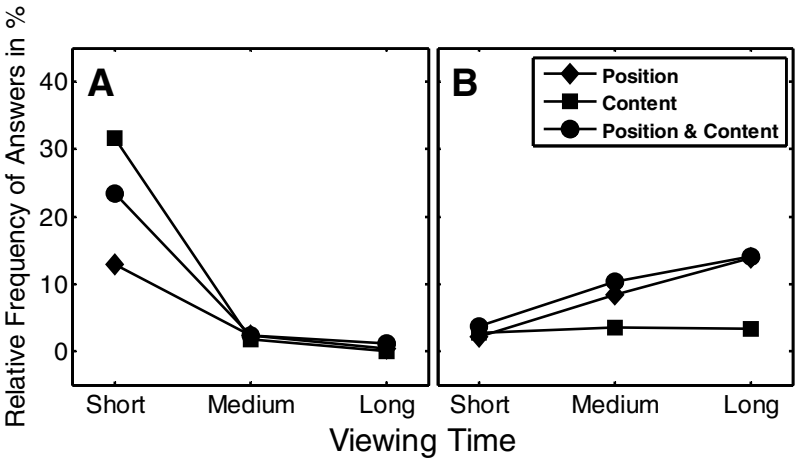


Fig. 2. Interaction of viewing time and task for (A) *previous* and (B) *next* reactions

from *short* (22.6%) to *medium* (2.2%) to *long* (.5%). The interaction dramatically illustrates the specific relationship between viewing time and task condition for the probabilities (Figure 2A). Post hoc analyses reveal significant differences between tasks for short viewing time only, $F(2,34) = 6.08$, $p < .001$, $\eta^2 = .11$.

In the case of *next*, significant main effects were found for task $F(2, 34) = 5.302$, $p = .001$, $\eta^2 = .064$, and viewing time $F(2, 34) = 8.010$, $p < .001$, $\eta^2 = .086$, as well as the interaction of both factors, $F(4, 68) = 5.583$, $p < .001$, $\eta^2 = .037$. With regard to task, the lowest probability was obtained for *content* (3.1%), followed by *position* (8.1%) and *position & content* (9.3%). Looking at viewing time, probabilities increase with viewing time (*short* 2.8%, *medium* 7.4%, and *long* 10.4%). As already described for *previous* reactions, the interaction of both factors shows a systematic pattern (Figure 2B). Here, post hoc analyses revealed significant differences between tasks in the *medium*, $F(2,34) = 4.69$, $p = .016$, $\eta^2 = .084$, and *long* condition, $F(2,34) = 8.35$, $p = .001$, $\eta^2 = .15$, respectively.

3.3 Discussion

The objective of this experiment was to systematically analyze influences of different tasks on the report of the subjective 'last glance'. Our method permitted the performance of sequences of fixations and saccades—similar to naturalistic gaze behaviour—and furthermore contrasted different visual tasks, namely spatial localisation and identification.

The present results do not support the attention-ahead-of-fixation assumption. In fact, an asynchrony of actual eye position and reported position was only found for short viewing times, and the asynchrony was in direct contrast to studies where eye position was found to lag behind attention [23]. We observed a pronounced dependency of reports on the task at hand. The influence of the task on reporting behaviour also interacted with viewing time. We found a strong tendency to report the pictogram

from a previous position. In addition, we discovered a second trend: if the task explicitly involved localisation, the probability of reporting the subsequent fixation position grew with increasing viewing time. This second trend is akin to the results of Fischer [21] as well as to results of the large number of studies using spatial cueing paradigms [19]. However, the trend was by far weaker than that usually reported in the single-saccade spatial cueing experiments. According to our data, slightly more than 10% of fixations were reported as being shifted towards the next spatial location, even with the longest viewing time at an actual position.

4 Gaze Transfer in Remote Collaboration

The final section of this article is dedicated to another domain where eye movements are of central importance: social interaction. Several studies have investigated the role of gaze behaviour in direct social interaction [for review 24] and when interacting with a virtual character [25]. However, in contemporary work life, a major percentage of social interactions take place in the form of remote collaboration, with the partners residing in different locations [26]. It has been demonstrated that transferring the gaze of one partner to the other partner with a cursor superimposed on the visual material can improve the performance in spatial tasks by disambiguating object references [27].

Regardless of the benefits of gaze transfer compared to purely verbal interactions, transferring computer mouse positions provides a rather direct pointing device, also allowing for referential disambiguation. Despite the lack of performance differences between gaze and mouse transfer, both transfer methods have differential effects on the cooperation process. Recent research has revealed that transferring gaze resulted in difficulties in interpreting communicative intention but demonstrated a strong coupling of attention to the transferred cursor [28]. Consequently, gaze transfer required a more effortful verbal disambiguation.

Here, we investigated the effects of gaze and mouse cursor transfer under conditions where the information about a person's attention and search process was crucial. Due to the strong link between attention and eye movements, a gaze cursor could be expected to provide an advantage over purely intentional mouse pointing. Our goal was to determine how the usability of gaze or mouse cursor transfer depends on the partner's ability to link this cursor to the objects in question. Pairs of participants had to solve a joint path-selection task with a strong spatial component on different processing levels (colour differentiation, form identification, calculation).

4.1 Method

Subjects. Forty-eight subjects (32 females) with a mean age of 23.9 years (range 18-51 years) participated in the experiment. They were invited in pairs and assigned to one of two experimental roles (searcher or assistant), resulting in a total of 24 pairs. All subjects reported normal or corrected-to-normal vision and received either course

credit or a compensation of €7 for participation in the study conducted in conformity with the declaration of Helsinki.

Apparatus. Both participants were seated in front of their computers in the same room, separated by a portable wall. The computers were connected via Ethernet. Eye movements of the searcher were recorded monocularly at 500 Hz with the SR EyeLink 1000 infrared eye tracking system in the remote recording mode.

Stimuli. Twenty images with a resolution of 1024 by 768 pixels served as stimuli in the experiment. They were composed of a grid of 20 x 20 rectangles, forming three red and three green paths (see Figure 3B). On each path, a variable number of circles and triangles was positioned, containing positive or negative digits; this information was only visible to the searcher within a window of 255 x 190 pixels (1/16 of the screen), while the rest of the screen was covered in black (Figure 3A). For the assistant, the whole screen area was visible. In the condition *objects*, all objects were depicted as circles (Figure 3B), while in the condition *grid*, only a grey background consisting of equidistant vertical and horizontal lines was visible (Figure 3C). The searcher's gaze or mouse position was projected onto the assistant's screen as a tricolour eye-icon.

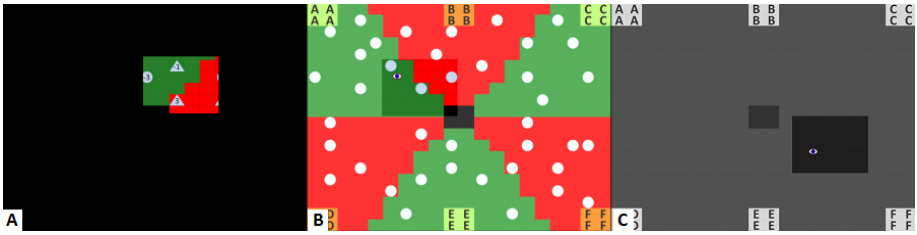


Fig. 3. Stimuli for searcher (A), and for assistant in the conditions objects (B) and grid (C)

Procedure. The experiment consisted of four blocks corresponding to the combinations of the experimental conditions (see below). The basic task in all experimental conditions was the following: In five trials per block, participants had to determine the correct path in a stepwise manner. They first had to select the three red paths, next they had to exclude the path with the least number of circles before finally determining the path which contained the smaller sum of digits. The chosen path had to be selected by the searcher via mouse click on the respective letter target field (see Figure 3B). The form of the paths was identical throughout the whole experiment, only their order changed across trials.

The searcher was provided with the full stimulus information necessary to solve the task, but saw only a section of the display (Figure 3A), while the assistant had to move this viewing window to reveal the respective display areas relevant to the searcher. Either the searcher's gaze or mouse cursor was transferred to the assistant for guidance. Besides the cursor, the assistant either saw the object positions, but not their identity (Figure 3B), or had no task-relevant visual information (Figure 3C). Participants were free to verbally interact in all experimental conditions. We recorded

the eye movements of the searcher, the mouse actions of both participants, and their verbal interactions.

4.2 Results

Performance. Mean solutions times were applied to a 2 (cursor: gaze, mouse) × 2 (assistant view: objects, grid) repeated measures ANOVA and revealed main effects for cursor, $F(1,23) = 22.60$, $p < .001$, $\eta^2 = .080$, and assistant view, $F(1,23) = 14.69$, $p < .001$, $\eta^2 = .102$, as well as a significant interaction, $F(1,23) = 11.89$, $p = .002$, $\eta^2 = .041$. Solution times were shorter in mouse than in gaze (*M*s: 78 vs. 102 s) and shorter in objects than in grid (*M*s: 76 vs. 103 s). The interaction results from the fact that the difference between mouse and gaze was only present in grid, $p < .001$, but not in objects, $p = .153$ (see Figure 4A). Mean error rates were 19.4% and showed no differences between the experimental conditions, all $F < 3$, all $p > .1$.

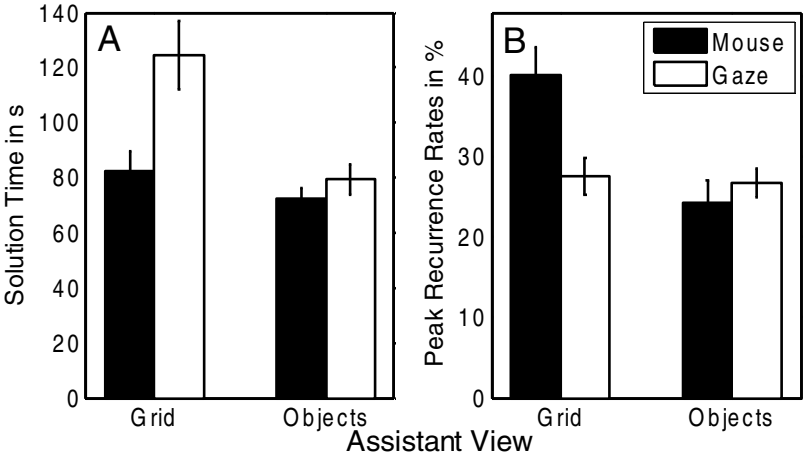


Fig. 4. Solution times (A) and peak recurrence rates between cursor and window centre (B) for the investigated cursor conditions.

Window and Cursor Alignment. In order to examine the positional coupling between the searcher's and the assistant's cursor (with the latter one corresponding to the window centre), a cross recurrence analysis [29] was conducted. This analysis provides cursor recurrence rates as the percentage of samples where the position of searcher and assistant cursor were located at about the same position, qualified by all cursor samples and at different temporal delays between both time series. Peak recurrence rates were subjected to a 2 (cursor: gaze, mouse) × 2 (assistant view: objects, grid) repeated measures ANOVA, revealing main effects for cursor, $F(1,23) = 5.16$, $p = .033$, $\eta^2 = .034$, and assistant view, $F(1,23) = 18.59$, $p < .001$, $\eta^2 = .089$, as well as an interaction, $F(1,23) = 19.78$, $p < .001$, $\eta^2 = .073$. The coupling was stronger in mouse than in gaze (32.3 vs. 27.2%) and stronger in grid than in objects (33.9 vs. 25.6%). However, grid recurrence rates were larger for

mouse than for gaze, $p < .001$, whereas no such difference was observed in objects, $p = .705$ (see Figure 4B).

In addition to peak recurrence, the temporal dynamics of the cursor window alignment were investigated. When considering recurrence rates as a curve ascending from a maximum negative temporal delay (-1000 ms in this study), peaking at a certain delay and then descending until maximum positive delay (1500 ms), the resulting amplitude of this curve represents a measure for the degree to which recurrence rates depend on temporal delay. Thus, it provides a measure of how tightly two cursors are coupled. There was an effect of cursor, $F(1,23) = 11.95$, $p = .002$, $\eta^2 = .043$, an effect of assistant view, $F(1,23) = 26.48$, $p < .001$, $\eta^2 = .084$, and an interaction between both factors, $F(1,23) = 59.00$, $p < .001$, $\eta^2 = .109$. When using the mouse, the increase was higher in grid than in objects (15.9 vs. 6.2%), $p < .001$. However, when gaze was used, there was no difference between grid and objects (7.5 and 8.1%), $p = .444$.

4.3 Discussion

Transferring gaze without further visual information about the task environment resulted in longer solution times compared with mouse transfer. Similar solution times were found for gaze and mouse transfer when adequate visual information was available. In this case, seeing a partner's gaze position can be as helpful as seeing their mouse. Without the required visual information, subjects were still able to efficiently use the mouse but not the gaze cursor. How can this effect be accounted for?

We think that the interpretability of the different cursors is directly related to the information they transmit. This information differs between gaze and mouse. Eye movements provide a rather direct visualization of visual attention in relation to task-relevant objects, especially in active tasks [30]. Their temporal and spatial parameters are closely related to processing information about these objects [11]. Visual attention usually does not float freely in space, but always implies a relation between a person and the entities that are being attended to. Thus, transmitting eye movements without an appropriate framework makes it difficult to interpret the gaze behaviour. In contrast, using the mouse as an intentional device for communication allows solely employing it to give messages to the partner. Thus, the partner knows that whatever the mouse does, he can simply react. In fact, several searchers instructed their assistants to "don't think, just follow my cursor". Although the assistants do not understand why a certain mouse movement is executed, they can be sure that it is produced as a deictic sign. In this case, simply following the cursor is a suitable strategy.

Our cross recurrence data support this interpretation. For mouse transfer, the coupling between the searcher's cursor and the window centre increased when no objects were available. Thus, the assistant relied more strongly on the searcher's guidance. Such an increased coupling was not found for gaze transfer. Additionally, the change of recurrence rate over different temporal delays was more than 2.5 times higher for grid than objects in the mouse condition but recurrence rate was similar for gaze in both viewing conditions. Thus, impoverished viewing conditions resulted in closer coupling of attention to mouse movements but not to gaze behaviour.

What is the benefit of mouse movements, and why can they be used more reliably than gaze, at least in the grid condition? While gaze is too fast and unpredictable to be followed unselectively, mouse movements can be adjusted to the requirements by performing slow and systematic moves. Thus, people can follow the mouse without necessarily understanding the meaning of the individual moves.

Since gaze transfer cannot be controlled as easily by the user, future research will focus on finding ways of adjusting gaze so that it will provide appropriate support. More refined visualization techniques appear to be a fruitful approach to address this issue. For example, smoothing the transferred gaze positions has been shown to increase subjective cursor control in a human computer interaction setting [31] and might improve its usability in cooperative situations as well.

5 General Discussion

The current work addressed three important issues regarding the analysis and interpretation of eye movements in the context of cognitive research and application.

In the first investigation, we demonstrated the susceptibility of visual fixation behaviour to the appearance of irrelevant visual or auditory distractors. According to our results, fixations are sensitive to multimodal distractions throughout their complete time course. Two main conclusions can be drawn from this study. First, when using fixation duration as a measure of information processing, a careful consideration of ongoing activities in the environment is required in order to avoid confounding artefacts in the measurement. Second, this finding should not be understood as a general rejection of the use of fixation durations. Rather we want to emphasize that using this parameter can provide helpful insights into ongoing processing mechanisms. To elaborate on this, Velichkovsky and colleagues [15] took advantage of the high sensitivity of fixations to sudden changes in the environment by suggesting a model for hazard detection, based on the instantaneous increase of fixation durations. Furthermore, it has been suggested that this particular feature of visual fixations can be used as a probe, providing access to different processing modes simply by a systematic presentation of such distractors [32].

The second experiment was concerned with the focus of visual attention by differentiating between the physical location of the eye and the subjective impression of what has been sampled from a visual display. Two main findings can be identified. A correspondence between the direction of the eye and the allocation of attention cannot always be assumed. In contrast to previous studies [19], our analysis is based on a continuous visual task, which can be understood as a viewing task close to natural gaze behaviour. The other key finding is related to the influence of the task. By differentiating between identification and localisation, we discovered a systematic relationship between the position of the eye on the display and the allocation of visual attention. More precisely, we found that in a task that requires identification—which is usually a more demanding and slower process than localisation—there is a higher probability that processing lags behind the actual position of the eye, whereas the

opposite is the case when only localisation is required. Thus, when employing gaze behaviour in the dialogue with technical devices, for instance in the context of attentive interfaces, a careful consideration of the task is required. Inferring the allocation of attention solely on the basis of gaze direction can be misleading because visual attention can be ahead or behind the position of the eye. Even if these mismatches in synchronization might have a magnitude of several milliseconds only, they need to be considered for an optimal design of such attentive interfaces.

The final study investigated the contribution of gaze in interactive settings where direct communication is impaired due to spatial separation. Employing gaze transfer in remote collaboration has been a question of interest for research as well as for application. Research questions about gaze transfer are mainly concerned with the problem of how much information can be provided by transferring eye movements and what inferences are possible on the side of the receiver [28]. The inability to find advantages of attention transfer over purely intentional forms of spatial referencing poses serious questions about the information required in the process of establishing a shared understanding [33]. From a more applied perspective, the efficient transfer of knowledge (i.e. expertise) across long distances is of particular importance. The conclusion from the present experiment is—similar to our second study—that task characteristics have to be taken into account when applying gaze transfer. It is possible to follow a mouse cursor almost blindly but when using gaze cursors it is of paramount importance that the recipient perceives them in relation to the corresponding environment. This evokes two further questions: When applying gaze transfer in a more complex task than the one we used here, for instance when specific skills or expertise have to be communicated, which cursors (gaze or mouse) would allow for a better knowledge transmission? It might be the case that seeing someone's gaze behaviour always provokes at least a minimum of interpretation activity, compared with a more mechanistic following of the mouse movement. Thus, the first case is preferable when knowledge transfer is intended. Another important issue is the workload on the side of the expert/transmitter. It can be expected that gaze transfer is less of a burden than requiring explicit mouse movements all the time. Finally, one can even think of reversing the whole paradigm: Why not transfer the novice's gaze to the expert? Making the gaze behaviour of the novice available should allow the expert to easily identify critical instances where support is required.

On the basis of three concrete examples, the research presented here illustrates the potential contribution of human eye movements to understanding mechanisms of information processing. Together with the advantages, we highlighted possible pitfalls in the interpretation and application of eye gaze analysis. Only a careful implementation—considering the many facets of gaze behavior—will allow using the potential of eye movements when developing cognitive technical systems.

Acknowledgments. We are grateful to Cathy Nangini for valuable suggestions. This research was supported by the FP7-PEOPLE-2009-IEF program #254638 to SP, a grant of the TU Dresden Centre for Continuing Education to RM, and the Russian Foundation for Basic Research (#09-06-12003) to BMV.

References

1. Cavanagh, P.: Attention routines and the architecture of selection. In: Posner, M.I. (ed.) Cognitive Neuroscience of Attention, pp. 13–28. Guilford Press, New York (2004)
2. Henderson, J.M.: Regarding scenes. Current Directions in Psychological Science 16, 219–222 (2007)
3. Land, M.F.: Eye movements and the control of actions in everyday life. Progress in Retinal and Eye Research 25, 296–324 (2006)
4. Findlay, J.M.: Active vision: Visual activity in everday life. Current Biology 8, R640–R642 (1998)
5. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active Vision. International Journal of Computer Vision 1, 333–356 (1987)
6. Matin, E.: Saccadic suppression: A review and an analysis. Psychological Bulletin 81, 899–917 (1974)
7. Kowler, E.: Eye movements: the past 25 years. Vision Research 51, 1457–1483 (2011)
8. Rayner, K.: Eye movements and attention in reading, scene perception, and visual search. The Quarterly Journal of Experimental Psychology 62, 1457–1506 (2009)
9. Vertegaal, R., Shell, J.S., Chen, D., Mamuji, A.: Designing for augmented attention: Towards a framework for attentive user interfaces. Computers in Human Behavior 22, 771–789 (2006)
10. Senders, J.W.: Four theoretical and practical questions. Keynote address presented at the Eye Tracking Research and Applications Symposium 2000. In: Duchowski, A.T. (ed.) Proceedings of the Symposium on Eye Tracking Research & Applications, p. 8. ACM Press, Palm Beach Gardens (2000)
11. Velichkovsky, B.M.: Heterarchy of cognition: The depths and the highs of a framework for memory research. Memory 10, 405–419 (2002)
12. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124, 372–422 (1998)
13. Castelhano, M.S., Mack, M.L., Henderson, J.M.: Viewing task influences eye movement control during active scene perception. Journal of Vision 9(6), 1–15 (2009)
14. Reingold, E.M., Stampe, D.M.: Saccadic inhibition and gaze contingent research paradigms. In: Kennedy, A., Radach, R., Heller, D., Pynte, J. (eds.) Reading as a Perceptual Process, pp. 1–26. Elsevier Science Ltd. (2000)
15. Velichkovsky, B.M., Rothert, A., Kopf, M., Dornhoefer, S.M., Joos, M.: Towards an express diagnostics for level of processing and hazard perception. Transportation Research, Part F 5, 145–156 (2002)
16. Pannasch, S., Velichkovsky, B.M.: Distractor effect and saccade amplitudes: Further evidence on different modes of processing in free exploration of visual images. Visual Cognition 17, 1109–1131 (2009)
17. Levine, T.R., Hullett, C.: Eta-square, partial eta-square, and misreporting of effect size in communication research. Human Communication Research 28, 612–625 (2002)
18. Pannasch, S., Dornhoefer, S.M., Unema, P.J.A., Velichkovsky, B.M.: The omnipresent prolongation of visual fixations: Saccades are inhibited by changes in situation and in subject's activity. Vision Research 41, 3345–3351 (2001)
19. Schneider, W.X., Deubel, H.: Selection-for-perception and selection-for-spatial-motor-action are coupled by visual attention. In: Prinz, W., Hommel, B. (eds.) Attention and Performance XIX: Common Mechanisms in Perception and Action, pp. 609–627. Oxford University Press, Oxford (2002)

20. Tatler, B.W.: Characterising the visual buffer: real-world evidence for overwriting early in each fixation. Perception 30, 993–1006 (2001)
21. Fischer, M.H.: An investigation of attention allocation during sequential eye movement tasks. The Quarterly Journal of Experimental Psychology: Human Experimental Psychology 3, 649–677 (1999)
22. Carr, T.H.: A multilevel approach to visual attention. In: Posner, M.I. (ed.) Cognitive Neuroscience of Attention, pp. 56–70. The Guilford Press, New York (2004)
23. Deubel, H., Irwin, D.E., Schneider, W.X.: The subjective direction of gaze shifts long before the saccade. In: Becker, W., Deubel, H., Mergner, T. (eds.) Current Oculomotor Research: Physiological and Psychological Aspects, pp. 65–70. Plenum, New York (1999)
24. Frischen, A., Bayliss, A.P., Tipper, S.P.: Gaze Cueing of Attention: Visual Attention, Social Cognition, and Individual Differences. Psychological Bulletin 133, 694–724 (2007)
25. Schrammel, F., Pannasch, S., Graupner, S.-T., Mojzisch, A., Velichkovsky, B.M.: Virtual friend or threat? The effects of facial expression and gaze interaction on psychophysiological responses and emotional experience. Psychophysiology 46, 922–931 (2009)
26. Hill, E.J., Ferris, M., Märtinson, V.: Does it matter where you work? A comparison of how three work venues (traditional office, virtual office, and home office) influence aspects of work and personal/family life. Journal of Vocational Behavior 63, 220–241 (2003)
27. Velichkovsky, B.M.: Communicating attention: Gaze position transfer in cooperative problem solving. Pragmatics and Cognition 3, 199–222 (1995)
28. Mueller, R., Helmert, J.R., Pannasch, S., Velichkovsky, B.M.: Improving remote cooperation in spatial tasks: Benefits and pitfalls of the gaze transfer approach. Quarterly Journal of Experimental Psychology (in revision)
29. Marwan, N., Kurths, J.: Nonlinear analysis of bivariate data with cross recurrence plots. Physics Letters A 302, 299–307 (2002)
30. Land, M.F., Tatler, B.W.: Looking and Acting: Vision and Eye Movements during Natural Behaviour. Oxford University Press (2009)
31. Helmert, J.R., Pannasch, S., Velichkovsky, B.M.: Influences of dwell time and cursor control on the performance in gaze driven typing. Journal of Eye Movement Research 2(3), 1–8 (2008)
32. Pannasch, S., Schulz, J., Velichkovsky, B.M.: On the control of visual fixation durations in free viewing of complex images. Attention, Perception, & Psychophysics 73, 1120–1132 (2011)
33. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (eds.) Perspectives on socially shared cognition, pp. 127–149. APA Books, Washington, DC (1991)

Ten Recent Trends in Computational Paralinguistics

Björn Schuller and Felix Weninger

Institute for Human-Machine Communication, Technische Universität München, Arcisstr. 21, 80333 München, Germany
{schuller,weninger}@tum.de

Abstract. The field of computational paralinguistics is currently emerging from loosely connected research on speaker states, traits, and vocal behaviour. Starting from a broad perspective on the state-of-the-art in this field, we combine these facts with a bit of 'tea leaf reading' to identify ten currently dominant trends that might also characterise the next decade of research: taking into account more tasks and task interdependencies, modelling paralinguistic information in the continuous domain, agglomerating and evaluating on large amounts of heterogeneous data, exploiting more and more types of features, fusing linguistic and non-linguistic phenomena, devoting more effort to optimisation of the machine learning aspects, standardising the whole processing chain, addressing robustness and security of systems, proceeding to evaluation in real-life conditions, and finally overcoming cross-language and cross-cultural barriers. We expect that following these trends we will see an increase in the 'social competence' of tomorrow's speech and language processing systems.

Keywords: Computational paralinguistics, speech analysis, speaker classification, machine learning.

1 Introduction

Social competence, i. e., the ability to permanently analyse and re-assess dialogue partners with respect to their traits (e. g., personality or age) and states (e. g., emotion or sleepiness), and to react accordingly (by adjusting the discourse strategy, or aligning to the dialogue partner) remains one key feature of human communication that is not found in most of today's technical systems. By simulating such capabilities through signal processing and machine learning techniques, the emerging field of computational paralinguistics aims to increase the perceived social competence of technical systems for human-machine communication. One main application is to increase efficiency and hence, user satisfaction in task oriented dialogue systems by enabling naturalistic interaction. Furthermore, recognition of paralinguistic information in human signals can be used for multimedia retrieval (enabling queries by certain speaker traits), in surveillance applications (e. g., to monitor customer satisfaction or potential attackers), for efficient audio or video coding and speech-to-speech translation (e. g., resolving semantic ambiguities by recognising

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 35–49, 2012.

intention, or synthesising translated speech with the original speaker's affect) and finally entertainment (e. g., to render states and traits in the voice of an avatar in accordance to the player).

As can be seen from these applications, computational paralinguistics comprise a variety of tasks [84]. A taxonomy can be established along the time axis, distinguishing long term *traits* from medium-term phenomena and short-term *states*. Long term traits include biological primitives such as height, weight, age, gender or race [55,77,81]. Interestingly, humans seem to exploit acoustic correlates of these primitives in their reasoning. For instance, age, height, and weight of speakers could be assigned by listeners to voices in repeated studies [22,42]; acoustic correlates of body shape, size and weight include fundamental frequencies and other formant parameters [28,35]. Other trait concepts such as group membership, ethnicity and culture overlap with linguistic phenomena such as dialect or nativeness [57]. For instance, the output of a speech recognition system can be used for classification of demographic traits including education level, ethnicity, and geographic region [33]. Besides, analysis of personality is an increasingly popular area of research [34,53,59] comprising acoustic and linguistic phenomena [65]. Medium term speaker attributes refer to temporary conditions, including sleepiness [41], (alcohol) intoxication [45,68,78], health [50] or depression [25], but also group roles [43], friendship and identity [38]. Finally, important short term states from an application point of view include voice quality, speaking style, and affect. In typical applications, one will rarely encounter full-blown, prototypical emotions such as sadness or disgust, but rather affect-related states including interest [98], uncertainty [47], frustration [2], stress level [37] or pain [5].

In summary, we hope that this unified view on the aspects of computational paralinguistics in speech may help to bridge the gap between some of the loosely connected fields in speech processing, including speech and speaker recognition, and the emerging domain of speaker classification. Further, this view enables us to outline *ten trends* that might characterise the field of computational paralinguistics in the following years. These trends are partially motivated by technological development—first and foremost, drastic decreases in the cost of computing power and storage space, the latter enabling access to virtually infinite amounts of speech data—but also conceptual advances in machine learning and signal processing. Altogether, we believe, these will allow technologies for computational paralinguistics to penetrate into daily life, which poses, in turn, several 'grand challenges' connected to real-life applications as opposed to 'in-the-lab' usage. While we put a strong focus on speech *analysis* in this chapter, many of the trends might be relevant for speech *synthesis* as well.

2 Ten Recent and Future Trends

2.1 More Tasks and Coupling of Tasks

Relevant tasks in computational paralinguistics are manifold, and we have mentioned a non-exhaustive list of relevant tasks above. Still, the lion's share of research is devoted to emotion and emotion-related states, followed by physical

traits (age, height) and personality[1]. It can be conjectured that addressing additional tasks will largely depend on the availability of annotated data. However, it could turn out that taking into account more and more seemingly novel tasks would be reinventing the wheel: A number of inderdependencies is already visible in the above list of paralinguistic states and traits. Following the taxonomy along the time axis, many dependencies on long term traits can be found. Long term traits themselves are coupled to some degree, e. g., height with age, gender and race. Medium term phenomena can depend on long term traits as well, e. g., health state can deteriorate with age, and group roles arguably depend on personality traits such as leadership emergence. Finally, also short term states are dependent on long term traits: The manifestation of emotion is dependent on personality [62,63]; in [54], it was revealed that human listeners consistently associate different tones of voice with certain speaker personalities. Furthermore, gender-dependencies of non-linguistic vocalisations have been repeatedly reported, e. g., in [60] for laughter.

Indeed, it has been repeatedly confirmed that modelling 'contextual' knowledge from different paralinguistic tasks benefits the performance in practice. Such knowledge can be integrated by building age, gender or height dependent models for any of the other tasks. For example, several studies indicate that considering gender information enables higher accuracy of automatic speech emotion recognition [93,95]; however, it is an open question whether this can be attributed to low-level acoustic differences in the pitch registers of male and female voices, or to higher-level differences in the expression of emotion. To exploit mutual information from the speaker identity, speaker adaptation or normalisation can be performed [9]. Finally, related state and trait information can be added as a feature: In [81] first beneficial effects are shown by providing knowledge on speaker dialect region, education level and race as ground truth along with acoustic features in the assessment of speaker traits including age, gender and height. Such addition of speaker traits as ground truth features can be relevant in practical situations, where for example height can be determined from camera recordings.

An alternative to such explicit modelling of dependencies using prior knowledge is to automatically learn them from training data. For example, the rather simple strategy of using pairs of age and gender classes as learning target instead of each attribute individually can already be advantageous, as has been proven in the first Paralinguistic Challenge [77]. In the future, enhanced modelling of multiple correlated target variables could be performed through multitask learning [15]. Here, a representation of the input features is shared among tasks, such as the internal activations in the hidden layer of a neural network. Recurrent neural networks, in particular, allow accessing past predictions for any of the variables for analysing the current time frame [89]—in some sense, this is similar to replacing the ground truth speaker features in the above setup by (time-varying) classifier predictions. In this context, one of the peculiarities of computational paralinguistics is found in the representation of task variables

[1] According to a Scopus search for the title ('speech' or 'speaker') AND . . . in February 2011.

by various data types (continuous, ordinal, nominal), which additionally often differ by their time scale (e. g., gender is constant in a speech turn while emotion or speaking style may vary). Considering methods for multi-scale fusion, one could also exploit multi-task learning for integrating phoneme recognition with analysis of paralinguistic information, in order to increase robustness of conversational speech recognition. Coupling the speech recognition task with, for example, gender or dialect recognition could be beneficial since in [10] it was shown that both these traits affect speech rate, flapping and central vowels.

2.2 More Continuous Modelling

The classic approach to computational paralinguistics is classification into 2-n classes, e. g., the big 6 emotions, gender, or age groups [77]. However, this often corresponds to an artificial discretisation, implying loss of information. For instance, the ground truth is continuous in case of intoxication (blood or breath alcohol concentration) or physical speaker traits (age, weight and height). Concepts to measure emotion and personality are often based on continuous valued dimensions, of which the most common are the arousal-valence model [66], or the five-factor 'OCEAN' model of personality [20]. For some states, annotation is performed using ordinal scales, e. g., using the Karolinska Sleepiness Scale (KSS), resulting in a quasi-continuum when ratings from multiple annotators are fused, e. g., by averaging; emotion annotation is sometimes performed directly in continuous dimensions, e. g., by the Feeltrace toolkit [19]. Conversely, machine learning research provides a rich set of tools for predicting continuous quantities including (extensions of) logistic regression, support vector regression, (recurrent) neural networks or random forests (ensembles of regression trees) which can be applied to paralinguistic analysis [77,98]. Evaluation procedures for regression are readily available as well, and include correlation (Pearson), rank-correlation (Spearman) and determination coefficients (R^2), mean absolute or (root) mean squared error. In addition to continuous valued annotation, short-term variations of speaker states can be captured by a representing them as a function of time. For example, the Feeltrace toolkit [19] allows annotating emotion with a 'sampling frequency' of 10 ms. On the recognition side, this allows for dynamic classification or regression techniques, and investigation of diverse units of analysis including syllables, words or turns [97].

2.3 More, Synthesised, Agglomerated, and Cross Data

While it is a common belief in pattern recognition that there is 'no data like more data', publicly available speech data with rich annotation of paralinguistic information are still sparse. In fact, there are increasingly more databases ready for experimentation; the crux is that these often come with different labelling schemes (discrete, continuous, dimensional, categorical) and, in the context of speaker states, different strategies for elicitation (acted, induced, natural). This makes data agglomeration and evaluation across multiple corpora less straightforward than for other tasks, such as automatic speech recognition. On the other

hand, multi-corpus and cross-corpus evaluation, such as done in [56] for age and gender and recently in [79,91] for emotion, is crucial to assess generalisation of the models. In fact, experiments in cross-corpus emotion recognition suggest some overfitting to single corpora [79] which can only partly be alleviated by corpus or speaker normalisation. To make things worse, common techniques to reduce overfitting such as feature selection may exhibit low cross-data generalisation themselves [29]. Hence, acquiring more data for building robust and generalising emotion models can be seen as one of the great challenges for the future. Recent results show that combining different databases in a unified labelling scheme through data agglomeration or voting significantly improves performance [85]. Still, such unification of the labelling schemes introduces 'information loss'; late fusion techniques for multiple classifiers trained on single corpora using distinct labelling schemes could be an interesting direction for the future. In addition, the efficacy of semi-supervised learning[2] to leverage unlabelled speech data for emotion recognition has been repeatedly demonstrated [39,48,100,103]; yet, large-scale studies across multiple speaker states and traits, and using large amounts of data acquired from the web, are still to follow. Finally, a promising technique is synthesis of training data: In fact, it has been shown that generalisation properties of emotion models in a cross-corpus setting can be improved through joint training with both human and synthetic speech [72]. This result is very promising since synthetic speech can be easily produced in large quantities, and a variety of combinations of speaker states and traits can be simulated. It is hoped that this will yield good generalisation of models and facilitate learning of multiple tasks and their interdependencies (cf. above).

2.4 More and Novel Features

The features used in early research on speaker states and traits were motivated by the adjacent fields of automatic speech and speaker recognition. Thus, usage of spectral or cepstral features (Mel frequency cepstral coefficients, MFCCs) prevailed. In the meantime, a plethora of novel, mostly expert-crafted acoustic features, including perceptually motivated ones [49,78,99] or such that base on pre-classification [64] have been proposed and evaluated for paralinguistic analysis, along with the addition of more or less brute forced linguistic features (e. g., Bag of Words or Bag of N-grams). Furthermore, it has repeatedly been shown that enlarging the feature space can help boost accuracy [78,80]. An alternative direction is supervised generation of features through evolutionary algorithms [74] or unsupervised learning of features, e. g., through deep belief networks or sparse coding. Still, the challenge is less the efficient computation, or combination of features in more or less brute force approaches, but to systematically investigate the relations between different types of features, especially in

[2] In the context of automatic speech recognition, this is often referred to as *unsupervised* learning—we prefer the more common term *semi-supervised* to highlight the difference to purely unsupervised techniques such as clustering or latent semantic analysis.

terms of generalisation to cross-corpus or cross-task analyses: After all, it is not clear whether novel features indeed add new information, or observed increases in performance stem from (over-)fitting to specific data sets, acoustic conditions, speakers or content (such as in fixed language speaker state corpora).

2.5 More (Coupling of) Linguistics and Non-linguistics

Transmitting information through non-verbal channels is a crucial part of human-human communication. Besides the low-level acoustic manifestations of speaker states and traits, such non-verbal channels also include the use of non-linguistic vocalisations. Recently, there is renewed interest in the use of such vocalisations in computer-mediated human-human and human-machine communication [12,83]. Just as human communication uses both non-verbal and verbal expression, the ultimate solution will, of course, not be to define new research domains dealing only with non-verbal phenomena (Social Signal Processing [94]), or to differentiate between non-linguistic vocalisations alone (such as in [13]), but to attain joint access to the linguistic / non-linguistic channels by machines. On the analysis side, there are already a couple of studies on fusion of linguistic with non-linguistic information. The simplest, yet effective and efficient strategy is to integrate non-linguistic vocalisations as word-like entities into the linguistic string [83,98]; in contrast, a late fusion approach has been investigated in [32].

2.6 More Optimisation

With the increased maturity of computational paralinguistics, and an established basic methodology, more and more efforts are devoted to optimisation of the whole processing chain. First, the systematic optimisation of machine learning algorithms including feature selection, reduction and classification is facilitated through the increasing availability of public corpora with well-defined partitioning into training and test sets, such as the ones used for the first paralinguistic challenges [76–78]. More precisely, such optimisation steps can involve 'global' as well as 'local' feature selection for sub-sets of classes in hierarchical classification [44] or for different sub-units of speech [7]. Additionally, more and more optimisations are applied in classifier training, including balancing of skewed class distributions (e. g., by synthetic minority oversampling [16,76] or similar techniques), or instance selection, i. e., pruning of noisy training data or outliers [26,82]. The importance of selecting appropriate classifier parameters is well known in machine learning and consequently also for paralinguistic information retrieval, as reported, e. g., in [78]. Besides, there is an increasing trend towards fusion of multiple systems, as has been evident in the sequence of paralinguistic challenges [76–78]. Fusion can be applied to classifier decisions in hierarchical [44,101], hybrid [75] or ensemble architectures [73,86]; at an even higher level, fusing the output of entire recognition systems can successfully exploit their complementarity; for instance, majority voting among the systems from the best participants in the Interspeech 2009 Emotion Challenge yields the best result reported so far on the challenge corpus [71].

Apart from such general machine learning techniques, speech analysis provides specific starting points for optimisation, including speech clustering by emotional state for speaker identification [23,46] and speaker adaptation / normalisation, which is nowadays observed particularly for speaker state analysis [9]. Finally, also the process of capturing speech signals itself can be optimised, e. g., by using silent speech interfaces for stress detection and speaker verification [58].

2.7 More Standardisation

Arguably, the more mature and closer to real-life application the field of computational paralinguistic gets, the greater is the need for standardisation. Similarly as in the argument made in the previous section, standardisation efforts can be categorised along the signal processing chain. They include documentation and well-motivated grouping of features such as the CEICES Feature Coding Scheme [4], standardised feature sets as provided by the openSMILE [31] and openEAR [30] toolkits in this field, and machine learning frameworks such as the Weka environment [36]. Such Standardised feature extraction / classification allows to evaluate the feature extraction and classification components of a recognition system separately. To further increase the reproducibility and comparability of results, well-defined evaluation settings are needed, such as the ones provided by recurring 'challenge' events [76–78]. Finally, communication between system components in real-life applications requires standardisation of recognition results for dialogue management or speech synthesis, etc. This is currently achieved by markup languages for description of emotional states (EMMA [3], EmotionML [69], MIML [51]) or extensions of VoiceXML to model speaker states in dialogue systems.

2.8 More Robustness

Robustness issues in the context of paralinguistic analysis can be categorised into technical robustness on the one hand and security on the other hand. Technical robustness refers to robustness against signal distortions including additive noise, e. g., environmental noise or interfering speakers (cocktail party problem) and reverberation, but also artefacts of transmission due to package loss and coding. Many of these issues have been extensively studied in the context of automatic speech recognition, and a wealth of methods is available, including speech enhancement, robust feature extraction, model-based techniques (i. e., learning the distortions) and novel recognition architectures such as graphical models. On another level, the *security* of paralinguistic analysis systems pertains to recognising malicious mis-use, i. e., attempted fraud. Examples for fraud include feigning of age (e. g., in an audio-based system for parental control), degree of intoxication, or emotion (e. g., by faking anger in an automated voice portal system in order to be redirected to a human operator).

Still, the majority of research in computational paralinguistics assumes laboratory conditions, i. e., a direct connection to the recogniser via high-quality audio interfaces, and data is recorded from (often paid) volunteers instead of real users

with potentially malicious intentions. There do exist a few studies on technical robustness of affect analysis, e. g., [70,92,96]—other speaker classification tasks are yet to follow. Yet, studies on the security of computational paralinguistics are currently sparse; these include detection of fake emotions from facial expressions [102] and recognition of feigned depression and sleepiness [14,61]. This is in stark contrast to the efforts devoted to speaker verification, i. e., robustness of speaker recognition systems against feigning speaker identity [11]. Besides, little attention has been paid to the 'goats' of paralinguistic analysis: This is how non-malicious system users that systematically cause false alarms have been termed in the 'zoo' of speaker verification [21]. For instance, it is known that speaker identification is hindered by emotion [87], and personality analysis is influenced by the use of second language [18]. Future research should broaden this analysis to other influence factors such as tiredness or intoxication; multi-task learning of paralinguistic information could help systems to model these influences.

2.9 More Realism

Basically, there is agreement that in order to evaluate systems for paralinguistic analysis in conditions close to real-life application, realistic data are needed: That is, natural occurrences of states and traits such as sleepiness or personality, recorded in real-life acoustic environments and interaction scenarios, are required. Still, progress is slow; one of the reasons might be the high effort of collecting and annotating such data. Of course, the required type of data depends on the particular application; in many cases, realistic data corresponds to spontaneous and conversational, i. e., verbally unrestricted speech. Besides, realism concerns the choice of testing instances. In order to obtain a realistic estimate of system performance, these should not be restricted to prototypical, straightforward cases, such as ones with high human agreement [90]. If pre-selection is applied, e. g., to gain performance bounds, this should follow transparent, objective criteria instead of an 'intuitive' selection by experts. Realism further relates to pre-processing of data such as chunking according to acoustic, phonetic or linguistic criteria. Such chunking should either be oriented on low-level acoustic features (i. e., a voice activity based chunking, which can already be challenging in reverberant or noisy acoustic conditions). Alternatively, if linguistic or phonetic criteria are employed, these should be evaluated on speech recognition output, such as in [52], not forced alignment based on manual transliteration, such as in many of today's emotional corpora, e. g., [76]. If additional meta-information or common knowledge is exploited in the analysis process, this information should be obtained from publicly available sources, e. g., by web-based queries, rather than by including expert knowledge. Finally, real-life applications imply the requirement of speaker independence in most cases. This can be established by partitioning into train, development and test sets [76]; however, often cross-validation is employed especially in case of small data sets, in order to ensure significance of results and to avoid overfitting in case of small data sets. Using a three-fold speaker independent and stratified subdivision according to

simple criteria (e. g., splitting according to subject IDs) seems to be a reasonable compromise between transparency and statistical significance in that case.

2.10 More Cross-Cultural and Cross-Lingual Evaluation

One of the barriers to overcome if paralinguistic information retrieval systems are to be widely employed is to enable their use across cultural borders. Yet, cross-cultural effects make this task even more challenging. Concerning speech, it is still an open question which speaker states and traits manifest consistently across cultures and languages [8]. It seems intuitive that, for example, linguistic features used to express certain emotional states differ; yet, often one-to-one mappings between languages can be found. However, generally little attention is paid to the more subtle effects of the cultural background. Among others, the relative robustness of speaker identification to the language being spoken has been confirmed [6], resulting in performance differences that are small in magnitude, although they may be statistically significant [40]. Emotion recognition, on the other hand, has been shown to depend strongly on the language being spoken [17,27,88]; multimodal fusion might be a promising approach since some non-verbal behavioural signals, including laughter [67] or facial expressions [24] have been found to be largely independent of cultural background. Indeed, there is evidence that multimodality helps humans in cross-cultural emotion recognition [1]. In general, it might turn out that cross-cultural recognition of paralinguistic information is just another instance of learning correlated tasks: Recognising the race, ethnicity, dialect region, etc. of a person could help in determining his or her emotion state, but possibly even biological primitives such as age or gender. Thus, while the most obvious strategy to perform cross-cultural recognition is to build specifically adapted models, any of the other strategies discussed above in Section 2.1 could be promising as well.

3 Conclusions

Starting from a broad and unified overview of the field of computational paralinguistics, we outlined ten dominant trends that can be summarised as: extending the field to new and combined tasks and more variety in data, taking into account recent paradigms in machine learning, and moving from 'out-of-the-lab' to real-life application contexts. Despite these recent developments, there remain some 'black spots' in literature. These include the generalisation of features and models across paralinguistic information retrieval tasks; determination of meaningful confidence measures for paralinguistics in general, for instance, for use in dialogue systems; and finally, bridging the gap between analysis and synthesis of speaker states and traits, by transferring methodologies and the broader view on computational paralinguistics to enable multi-faceted speech synthesis, voice transformation, and benefit from it for (ad-hoc) training of analysis systems. Following these trends, we expect higher generalisation abilities of future systems for computational paralinguistics, and we look forward to experiencing their increasing application in real world contexts.

References

1. Abelin, A.: Cross-Cultural Multimodal Interpretation of Emotional Expressions - An Experimental Study of Spanish and Swedish. In: Proc. of Speech Prosody, ISCA (2004); no pagination
2. Ang, J., Dhillon, R., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. Interspeech, pp. 2037–2040. Denver (2002)
3. Baggia, P., Burnett, D.C., Carter, J., Dahl, D.A., McCobb, G., Raggett, D.: EMMA: Extensible MultiModal Annotation markup language (2007), http://www.w3.org/TR/emma/
4. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Amir, N.: Whodunnit – Searching for the Most Important Feature Types Signalling Emotional User States in Speech. Computer Speech and Language 25, 4–28 (2011)
5. Belin, P., Fillion-Bilodeau, S., Gosselin, F.: The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods 40(2), 531–539 (2008)
6. Bellegarda, J.R.: Language-independent speaker classification over a far-field microphone. In: Mueller, C. (ed.) Speaker Classification II: Selected Projects, pp. 104–115. Springer, Berlin (2007)
7. Bitouk, D., Verma, R., Nenkova, A.: Class-level spectral features for emotion recognition. Speech Communication 52(7-8), 613–625 (2011)
8. Boden, M.: Mind as Machine: A History of Cognitive Science, ch. 9. Oxford Univ. Press, New York (2008)
9. Bone, D., Black, M.P., Li, M., Metallinou, A., Lee, S., Narayanan, S.: Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors. In: Proc. of Interspeech, Florence, Italy, pp. 3217–3220 (2011)
10. Byrd, D.: Relations of sex and dialect to reduction. Speech Communication 15(1-2), 39–54 (1994)
11. Campbell, J.: Speaker recognition: a tutorial. Proceedings of the IEEE 85(9), 1437–1462 (1997)
12. Campbell, N.: On the use of nonverbal speech sounds in human communication. In: Proc. of COST 2102 Workshop, Vietri sul Mare, Italy, pp. 117–128 (2007)
13. Campbell, N., Kane, J., Moniz, H.: Processing 'yup!' and other short utterances in interactive speech. In: Proc. of ICASSP, Prague, Czech Republic, pp. 5832–5835 (2011)
14. Cannizzaro, M., Reilly, N., Snyder, P.J.: Speech content analysis in feigned depression. Journal of Psycholinguistic Research 33(4), 289–301 (2004)
15. Caruana, R.: Multitask learning: A knowledge-based source of inductive bias. Machine Learning 28, 41–75 (1997)
16. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
17. Chen, A.: Perception of paralinguistic intonational meaning in a second language. Language Learning 59(2), 367–409 (2009)
18. Chen, S.X., Bond, M.H.: Two languages, two personalities? examining language effects on the expression of personality in a bilingual context. Personality and Social Psychology Bulletin 36(11), 1514–1528 (2010)

19. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: Feeltrace: An instrument for recording perceived emotion in real time. In: Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, pp. 19–24 (2000)
20. Digman, J.M.: Personality Structure: emergence of the Five-Factor Model. Ann. Rev. Psychol. 41, 417–440 (1990)
21. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: Proc. of ICSLP (1998); no pagination
22. van Dommelen, W.A., Moxness, B.H.: Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. Language and Speech 38(3), 267–287 (1995)
23. Dongdong, L., Yingchun, Y.: Emotional speech clustering based robust speaker recognition system. In: Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP 2009, Tianjin, China, pp. 1–5 (2009)
24. Elfenbein, H., Mandal, M.K., Ambady, N., Harizuka, S.: Cross-Cultural Patterns in Emotion Recognition: Highlighting Design and Analytical Techniques. Emotion 2(1), 75–84 (2002)
25. Ellgring, H., Scherer, K.R.: Vocal Indicators of Mood change in Depression. Journal of Nonverbal Behavior 20, 83–110 (1996)
26. Erdem, C.E., Bozkurt, E., Erzin, E., Erdem, A.T.: RANSAC-based training data selection for emotion recognition from spontaneous speech. In: AFFINE 2010 - Proceedings of the 3rd ACM Workshop on Affective Interaction in Natural Environments, Co-located with ACM Multimedia 2010, Florence, Italy, pp. 9–14 (2010)
27. Esposito, A., Riviello, M.T.: The cross-modal and cross-cultural processing of affective information. In: Proceeding of the 2011 Conference on Neural Nets WIRN10: Proceedings of the 20th Italian Workshop on Neural Nets, vol. 226, pp. 301–310 (2011)
28. Evans, S., Neave, N., Wakelin, D.: Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. Biological Psychology 72(2), 160–163 (2006)
29. Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S.: Cross-Corpus Classification of Realistic Emotions Some Pilot Experiments. In: Proc. 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Valetta, pp. 77–82 (2010)
30. Eyben, F., Wöllmer, M., Schuller, B.: openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Proc. ACII, Amsterdam, pp. 576–581 (2009)
31. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. ACM Multimedia, Florence, Italy, pp. 1459–1462 (2010)
32. Eyben, F., Wöllmer, M., Valstar, M., Gunes, H., Schuller, B., Pantic, M.: String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In: Proc. 9th International IEEE Conference on Face and Gesture Recognition 2011 (FG 2011), Santa Barbara, CA, pp. 322–329 (2011)
33. Gillick, D.: Can conversational word usage be used to predict speaker demographics? In: Proc. of Interspeech, Makuhari, Japan, pp. 1381–1384 (2010)
34. Gocsál: Female listeners' personality attributions to male speakers: The role of acoustic parameters of speech. Pollack Periodica 4(3), 155–165 (2009)

35. Gonzalez, J.: Formant frequencies and body size of speaker: a weak relationship in adult humans. Journal of Phonetics 32(2), 277–287 (2004)
36. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 (2009)
37. Hansen, J., Bou-Ghazale, S.: Getting started with susas: A speech under simulated and actual stress database. In: Proc. EUROSPEECH 1997, Rhodes, Greece, vol. 4, pp. 1743–1746 (1997)
38. Ipgrave, J.: The language of friendship and identity: Children's communication choices in an interfaith exchange. British Journal of Religious Education 31(3), 213–225 (2009)
39. Jia, L., Chun, C., Jiajun, B., Mingyu, Y., Jianhua, T.: Speech emotion recognition using an enhanced co-training algorithm. In: Proceedings of the 2007 IEEE International Conference on Multimedia and Expo., ICME 2007, Beijing, China, pp. 999–1002 (2007)
40. Kleynhans, N.T., Barnard, E.: Language dependence in multilingual speaker verification. In: Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, Langebaan, South Africa, pp. 117–122 (November 2005)
41. Krajewski, J., Batliner, A., Golz, M.: Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. Behavior Research Methods 41, 795–804 (2009)
42. Krauss, R.M., Freyberg, R., Morsella, E.: Inferring speakers physical attributes from their voices. Journal of Experimental Social Psychology 38(6), 618–625 (2002)
43. Laskowski, K., Ostendorf, M., Schultz, T.: Modeling Vocal Interaction for Text-Independent Participant Characterization in Multi-Party Conversation. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, pp. 148–155 (2008)
44. Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. In: Proc. Interspeech, Brighton, pp. 320–323 (2009)
45. Levit, M., Huber, R., Batliner, A., Nöth, E.: Use of prosodic speech characteristics for automated detection of alcohol intoxination. In: Bacchiani, M., Hirschberg, J., Litman, D., Ostendorf, M. (eds.) Proc. of the Workshop on Prosody and Speech Recognition 2001, Red Bank, NJ, pp. 103–106 (2001)
46. Li, D., Wu, Z., Yang, Y.: Speaker recognition based on pitch-dependent affective speech clustering. Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence 22(1), 136–141 (2009)
47. Litman, D., Rotaru, M., Nicholas, G.: Classifying Turn-Level Uncertainty Using Word-Level Prosody. In: Proc. Interspeech, Brighton, UK, pp. 2003–2006 (2009)
48. Mahdhaoui, A., Chetouani, M.: A new approach for motherese detection using a semi-supervised algorithm. In: Machine Learning for Signal Processing XIX - Proceedings of the 2009 IEEE Signal Processing Society Workshop, MLSP 2009, pp. 1–6. IEEE, Grenoble (2009)
49. Mahdhaoui, A., Chetouani, M., Kessous, L.: Time-Frequency Features Extraction for Infant Directed Speech Discrimination. In: Solé-Casals, J., Zaiats, V. (eds.) NOLISP 2009. LNCS (LNAI), vol. 5933, pp. 120–127. Springer, Heidelberg (2010)
50. Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E.: PEAKS - A system for the automatic evaluation of voice and speech disorders. Speech Communication 51, 425–437 (2009)
51. Mao, X., Li, Z., Bao, H.: An Extension of MPML with Emotion Recognition Functions Attached. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 289–295. Springer, Heidelberg (2008)

52. Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., Steidl, S.: Emotion recognition using imperfect speech recognition. In: Proc. Interspeech 2010, Makuhari, Japan, pp. 478–481 (2011)
53. Mohammadi, G., Vinciarelli, A., Mortillaro, M.: The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions. In: Proc. SSPW 2010, Firenze, Italy, pp. 17–20 (2010)
54. Mokhtari, A., Campbell, N.: Speaking style variation and speaker personality. In: Proc. of Speech Prosody, Campinas, Brazil, pp. 601–604 (2008)
55. Mporas, I., Ganchev, T.: Estimation of unknown speakers' height from speech. International Journal of Speech Technology 12(4), 149–160 (2009)
56. Müller, C., Wittig, F., Baus, J.: Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs. In: Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, pp. 1305–1308 (2003)
57. Omar, M.K., Pelecanos, J.: A novel approach to detecting non-native speakers and their native language. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Dallas, Texas, pp. 4398–4401 (2010)
58. Patil, S.A., Hansen, J.H.L.: The physiological microphone (pmic): A competitive alternative for speaker assessment in stress detection and speaker verification. Speech Communication 52(4), 327–340 (2010)
59. Polzehl, T., Möller, S., Metze, F.: Automatically assessing personality from speech. In: Proceedings - 2010 IEEE 4th International Conference on Semantic Computing, ICSC 2010, Pittsburgh, PA, pp. 134–140 (2010)
60. Provine, R.: Laughter punctuates speech: linguistic, social and gender contexts of laughter. Ethology 15, 291–298 (1993)
61. Reilly, N., Cannizzaro, M.S., Harel, B.T., Snyder, P.J.: Feigned depression and feigned sleepiness: A voice acoustical analysis. Brain and Cognition 55(2), 383–386 (2004)
62. Reisenzein, R., Weber, H.: Personality and Emotion. In: Corr, P.J., Matthews, G. (eds.) The Cambridge Handbook of Personality Psychology, pp. 54–71. Cambridge University Press, Cambridge (2009)
63. Revelle, W., Scherer, K.: Personality and Emotion. In: Oxford Companion to the Affective Sciences, pp. 1–4. Oxford University Press, Oxford (2009)
64. Ringeval, F., Chetouani, M.: A vowel based approach for acted emotion recognition. In: INTERSPEECH 2008 - 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, pp. 2763–2766 (2008)
65. Rosenberg, A., Hirschberg, J.: Acoustic/Prosodic and Lexical Correlates of Charismatic Speech. In: Proc. of Interspeech, Lisbon, pp. 513–516 (2005)
66. Russel, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
67. Sauter, D.A., Eisner, F., Ekman, P., Scott, S.K.: Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc. of the National Academy of Sciences of the U.S.A. 107(6), 2408–2412 (2010)
68. Schiel, F., Heinrich, C.: Laying the foundation for in-car alcohol detection by speech. In: Proc. INTERSPEECH 2009, Brighton, UK, pp. 983–986 (2009)
69. Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I.: What Should a Generic Emotion Markup Language Be Able to Represent? In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 440–451. Springer, Heidelberg (2007)
70. Schuller, B.: Affective speaker state analysis in the presence of reverberation. International Journal of Speech Technology 14(2), 77–87 (2011)

71. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing (9/10), 1062–1087 (2011)
72. Schuller, B., Burkhardt, F.: Learning with Synthesized Speech for Automatic Emotion Recognition. In: Proc. 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, pp. 5150–5153 (2010)
73. Schuller, B., Jiménez Villar, R., Rigoll, G., Lang, M.: Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: Proc. ICASSP, Philadelphia, pp. I:325–I:328 (2005)
74. Schuller, B., Reiter, S., Rigoll, G.: Evolutionary feature generation in speech emotion recognition. In: Proc. Int. Conf. on Multimedia and Expo, ICME 2006, Toronto, Canada, pp. 5–8 (2006)
75. Schuller, B., Rigoll, G., Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proc. ICASSP, Montreal, pp. 577–580 (2004)
76. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 Emotion Challenge. In: Proceedings of 11th European Conference on Speech Communication and Technology, Interspeech 2009 – Eurospeech, Brighton, UK, September 6-10, pp. 312–315 (2009)
77. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The INTERSPEECH 2010 Paralinguistic Challenge – Age, Gender, and Affect. In: Proceedings of 11th International Conference on Spoken Language Processing, Interspeech 2010 – ICSLP, Makuhari, Japan, September 26-30, pp. 2794–2797 (2010)
78. Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J.: The Interspeech 2011 Speaker State Challenge. In: Proc. Interspeech, Florence, Italy, pp. 3201–3204 (2011)
79. Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Transactions on Affective Computing 1(2), 119–131 (2010)
80. Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., Rigoll, G.: Brute-Forcing Hierarchical Functionals for Paralinguistics: a Waste of Feature Space? In: Proc. ICASSP, Las Vegas, pp. 4501–4504 (2008)
81. Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G., Arsic, D.: Semantic Speech Tagging: Towards Combined Analysis of Speaker Traits. In: Proc. AES 42nd International Conference, Ilmenau, Germany, pp. 89–97 (2011)
82. Schuller, B., Zhang, Z., Weninger, F., Rigoll, G.: Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization. In: Proc. 2011 Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel (2011)
83. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.: Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior 27, 1760–1774 (2009)
84. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: Paralinguistics in Speech and Language—State-of-the-Art and the Challenge. Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language (2011) (to appear)
85. Schuller, B., Zhang, Z., Weninger, F., Rigoll, G.: Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? In: Proc. of INTERSPEECH, pp. 1553–1556. ISCA, Florence (2011)

86. Schwenker, F., Scherer, S., Schmidt, M., Schels, M., Glodek, M.: Multiple Classifier Systems for the Recogonition of Human Emotions. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 315–324. Springer, Heidelberg (2010)
87. Shahin, I.: Verifying speakers in emotional environments. In: IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2009, Ajman, UAE, pp. 328–333 (2009)
88. Shami, M., Verhelst, W.: Automatic classification of expressiveness in speech: A multi-corpus study. In: Mueller, C. (ed.) Speaker Classification II: Selected Projects, pp. 43–56. Springer, Berlin (2007)
89. Stadermann, J., Koska, W., Rigoll, G.: Multi-task learning strategies for a recurrent neural net in a hybrid tied-posteriors acoustic mode. In: Proc. of Interspeech 2005, pp. 2993–2996. ISCA, Lisbon (2005)
90. Steidl, S., Schuller, B., Batliner, A., Seppi, D.: The Hinterland of Emotions: Facing the Open-Microphone Challenge. In: Proc. ACII, Amsterdam, pp. 690–697 (2009)
91. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., Schuller, B.: Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. In: Proc. ICASSP, Prague, Czech Republic, pp. 5688–5691 (2011)
92. Tabatabaei, T.S., Krishnan, S.: Towards robust speech-based emotion recognition. In: Proc. IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, pp. 608–611 (2010)
93. Ververidis, D., Kotropoulos, C.: Automatic speech classification to five emotional states based on gender information. In: Proc. of 12th European Signal Processing Conference, Vienna, Austria, pp. 341–344 (2004)
94. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing 27, 1743–1759 (2009)
95. Vogt, T., Andre, E.: Improving automatic emotion recognition from speech via gender differentiation. In: Proc. of Language Resources and Evaluation Conference (LREC 2006), Genoa, Italy, pp. 1–4 (2006)
96. Weninger, F., Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognition of non-prototypical emotions in reverberated and noisy speech by nonnegative matrix factorization. Eurasip Journal on Advances in Signal Processing 2011(Article ID 838790), 16 pages (2011)
97. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE Journal of Selected Topics in Signal Processing 4(5), 867–881 (2010)
98. Wöllmer, M., Weninger, F., Eyben, F., Schuller, B.: Acoustic-Linguistic Recognition of Interest in Speech with Bottleneck-BLSTM Nets. In: Proc. of INTERSPEECH, Florence, Italy, pp. 77–80 (2011)
99. Wu, S., Falk, T.H., Chan, W.: Automatic speech emotion recognition using modulation spectral features. Speech Communication 53(5), 768–785 (2011)
100. Yamada, M., Sugiyama, M., Matsui, T.: Semi-supervised speaker identification under covariate shift. Signal Processing 90(8), 2353–2361 (2010)
101. Yoon, W., Park, K.: Building robust emotion recognition system on heterogeneous speech databases. In: Digest of Technical Papers - IEEE International Conference on Consumer Electronics, pp. 825–826 (2011)
102. Zhang, Z., Singh, V., Slowe, T., Tulyakov, S., Govindaraju, V.: Real-time Automatic Deceit Detection from Involuntary Facial Expressions. In: Proc. of CVPR, pp. 1–6 (2007)
103. Zhang, Z., Weninger, F., Wöllmer, M., Schuller, B.: Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition. In: Proc. Automatic Speech Recognition and Understanding Workshop (ASRU 2011). IEEE, Big Island (2011)

Conversational Speech Recognition in Non-stationary Reverberated Environments

Rudy Rotili[1], Emanuele Principi[1], Martin Wöllmer[2], Stefano Squartini[1], and Björn Schuller[2]

[1] Dipartimento di Ingegneria dell'Informazione
Università Politecnica delle Marche, Ancona, Italy
{r.rotili,e.principi,s.squartini}@univpm.it
[2] Institute for Human-Machine Communication
Technische Universität München, Germany
{woellmer,schuller}@tum.de

Abstract. This paper presents a conversational speech recognition system able to operate in non-stationary reverberated environments. The system is composed of a dereverberation front-end exploiting multiple distant microphones, and a speech recognition engine. The dereverberation front-end identifies a room impulse response by means of a blind channel identification stage based on the Unconstrained Normalized Multi-Channel Frequency Domain Least Mean Square algorithm. The dereverberation stage is based on the adaptive inverse filter theory and uses the identified responses to obtain a set of inverse filters which are then exploited to estimate the clean speech. The speech recognizer is based on tied-state cross-word triphone models and decodes features computed from the dereverberated speech signal. Experiments conducted on the Buckeye corpus of conversational speech report a relative word accuracy improvement of 17.48% in the stationary case and of 11.16% in the non-stationary one.

1 Introduction

In the recent years, several research efforts have been devoted to distant speech recognition (DSR) systems [14]. The motivation behind this is that DSR systems are perceived as more user-friendly, comfortable and intuitive than solutions using head-set microphones. The task still represents a great research challenge, as the acquired speech signal is more affected by distortions, such as noise and reverberation. In addition, if multiple speakers are present (e.g. in meetings), the presence of overlapping speech makes the task even more challenging.

In this paper, the focus is on DSR in reverberated environments, thus other causes of degradation will not be considered. According to [13], dereverberation techniques can be classified depending on the component of the DSR in which they operate. Signal-based approaches, in particular, dereverberate the microphone signals before the feature extraction stage. Here the attention is focused on these techniques, more specifically on the inverse filtering methods [6].

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 50–59, 2012.

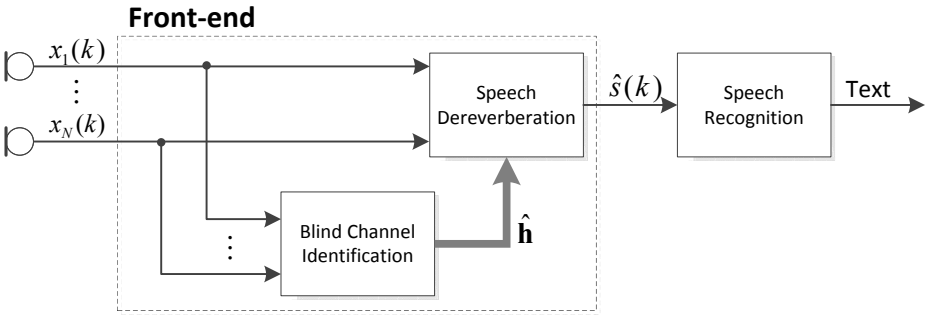


Fig. 1. System architecture

Assuming that the room impulse responses (RIRs) are available or estimated, inverse filtering methods aim at calculating a set of inverse filters for the RIRs and using them to dereverberate the microphone signals. Other signal-based approaches proposed in the literature cope with the reverberation problem by using beamforming techniques [14], spectral enhancement [6], non-negative matrix factorization [4], or linear-prediction residual enhancement [6].

This work proposes a system able to recognize conversational speech in a non-stationary reverberated acoustic environment (Fig. 1). The system is composed of a dereverberation front-end and a speech recognition engine which is based on the Hidden Markov Model toolkit (HTK) [17] and has been used as a baseline system in [15]. The front-end operates before the feature extraction stage and is based on the dereverberation algorithm proposed in [10] by some of the authors and on the identification algorithm proposed in [3]. It blindly identifies RIRs by means of the Unconstrained Normalized Multi-Channel Frequency Domain Least Mean Square (UNMCFLMS) algorithm, which are then equalized to recover the clean speech source. The recognizer processes cepstral mean normalized Mel-Frequency Cepstral Coefficient (MFCC) features and consists of context-dependent tied state cross-word triphone Hidden Markov Models (HMM) trained on conversational speech. A set of experiments have been conducted in stationary and non-stationary reverberated scenarios. The front-end capabilities of operating in non-stationary conditions have been assessed evaluating the *Normalized Projection Misalignment* (NPM) curves. The entire system has been evaluated in terms of word recognition accuracy on the artificially reverberated Buckeye corpus of conversational speech: The relative improvement over reverberated signals is 17.48% in the stationary case and 11.16% in the non-stationary one.

The outline of the paper is the following: Section 2 illustrates the blind dereverberation algorithm; Section 3 presents the conversational speech recognition system; Section 4 details the performed experiments and shows the obtained results; finally, Section 5 concludes the paper and presents some future developments.

2 Blind Dereverberation

2.1 Problem Statement

Let us consider a reverberant room with a single speech source and an array of N microphones, i.e. single-input multiple-output (SIMO) system. The observed signal at each sensor is then given by

$$x_n(k) = \mathbf{h}_n^T \mathbf{s}(k) \qquad n = 1, 2, \ldots, N \tag{1}$$

where $\mathbf{h}_n = [h_{n,0}\ h_{n,1} \cdots h_{n,L_h-1}]^T$ is the L_h-tap room impulse response between the source and n-th sensor, $\mathbf{s}(k) = [s(k)\ s(k-1) \cdots s(k-L_h+1)]^T$ is the input vector, and $(\cdot)^T$ denotes the transpose operator. Applying the z transform, equation (1) can be rewritten as:

$$X_n(z) = H_n(z)S(z), \qquad n = 1, 2, \ldots, N. \tag{2}$$

The objective is to obtain an estimate $\hat{s}(k)$ of the clean speech source by using only the microphone signals.

2.2 Blind Channel Identification

Considering the previously described SIMO system a Blind Channel Identification (BCI) algorithm aims to find the RIRs vector $\mathbf{h}_n$ by using only the microphone signals $x_n(k)$. Here, BCI is performed through the Unconstrained Normalized Multi-Channel Frequency-Domain Least Mean Square (UNMCFLMS) algorithm [3], an adaptive technique that offers a good compromise among fast convergence, adaptivity, and low computational complexity.

A brief review of UNMCFLMS now follows, please refer to [3] for details. The derivation of UNMCFLMS is based on cross relation criteria using the overlap and save technique. The frequency-domain cost function for the q-th frame is defined as

$$J_f = \sum_{n=1}^{N-1} \sum_{i=i+1}^{N} \underline{\mathbf{e}}_{ni}^H(q) \underline{\mathbf{e}}_{ni}(q) \tag{3}$$

where $\underline{\mathbf{e}}_{ni}(q)$ is the frequency-domain block error signal between the n-th and i-th channels and $(\cdot)^H$ denotes the Hermitian transpose operator. Defining $\mathbf{h}_{nm^*} = [\mathbf{h}_{1m^*}^T\ \mathbf{h}_{2m^*}^T\ \cdots \mathbf{h}_{Nm^*}^T\ ]^T$, the update equation of the UNMCFLMS is

$$\begin{aligned} \widehat{\underline{\mathbf{h}}}_{nm^*}(q+1) = & \widehat{\underline{\mathbf{h}}}_{nm^*}(q) - \rho[\mathbf{P}_{nm^*}(q) + \delta \mathbf{I}_{2L_h \times L_h}]^{-1} \\ & \times \sum_{n=1}^{N} \mathbf{D}_{x_n}^H(q) \underline{\mathbf{e}}_{ni}(q), \quad i = 1, 2, \ldots, N \end{aligned} \tag{4}$$

where $0 < \rho < 2$ is the step-size, δ is a small positive number and

$$\widehat{\underline{\mathbf{h}}}_{nm^*}(q) = \mathbf{F}_{2L_h \times 2L_h} \left[\widehat{\mathbf{h}}_{nm^*}(q)\ \mathbf{0}_{1 \times L_h} \right]^T \tag{5}$$

$$\underline{\mathbf{e}}_{ni}(q) = \mathbf{F}_{2L_h \times 2L_h} \left[\mathbf{0}_{1\times L_h} \ \left\{ \mathbf{F}^{-1}_{L_h \times L_h} \underline{\mathbf{e}}_{ni}(q) \right\}^T \right]^T \tag{6}$$

$$\mathbf{P}_{nm^*}(q) = \sum_{n=1, n\neq i}^{N} \mathbf{D}^H_{x_n}(q) \mathbf{D}_{x_n}(q). \tag{7}$$

$\mathbf{F}$ denotes the Discrete Fourier Transform (DFT) matrix. The frequency-domain error function $\underline{\mathbf{e}}_{ni}(q)$ is given by

$$\underline{\mathbf{e}}_{ni}(q) = \mathbf{D}_{x_n}(q) \widehat{\underline{\mathbf{h}}}_{nm^*}(q) - \mathbf{D}_{x_i}(q) \widehat{\underline{\mathbf{h}}}_{im^*}(q) \tag{8}$$

where the diagonal matrix

$$\mathbf{D}_{x_n}(q) = \mathrm{diag}\left(\mathbf{F} \left\{ [x_n(qL_h - L_h) \ x_n(qL_h - L_h + 1) \ \cdots \ x_n(qL_h + L_h - 1)]^T \right\} \right) \tag{9}$$

is the DFT of the q-th frame input signal block for the n-th channel. In order to guarantee proper convergence and a non-zero error signal, the algorithm is initialized in the time domain to satisfy the unit-norm constraint:

$$\widehat{\mathbf{h}}_n(0) = \left[1/\sqrt{N} \ 0 \ \cdots \ 0 \right]^T, \qquad n = 1, 2, \ldots, N. \tag{10}$$

From a computational point of view, the UNMCFLMS algorithm ensures an efficient execution of the circular convolution by means of the Fast Fourier Transform (FFT). In addition, it can be easily implemented for a real-time application since the normalization matrix $\mathbf{P}_{nm^*}(q) + \delta \mathbf{I}_{2L_h \times L_h}$ is diagonal, and it is straightforward to compute its inverse.

Though UNMCFLMS allows the estimation of long RIRs, it requires a high input signal-to-noise ratio. In this paper, the presence of noise has not been taken into account, therefore the UNMCFLMS is an appropriate choice, but different solutions have been proposed in literature in order to alleviate the problem [6].

2.3 Adaptive Inverse Filtering

Given the N room transfer functions (RTFs) $H_n(z)$, a set of inverse filters $G_n(z)$ can be found by using the Multiple-Input/Output Inverse Theorem (MINT) [5] such that

$$\sum_{n=1}^{N} H_n(z) G_n(z) = 1, \tag{11}$$

assuming that the RTFs do not have any common zeros. In the time-domain, the inverse filter vector denoted as $\mathbf{g}$, is calculated by minimizing the following cost function:

$$C = \left\| \mathbf{H}\mathbf{g} - \mathbf{v} \right\|^2, \tag{12}$$

where $\| \cdot \|$ denote the l_2-norm operator and $\mathbf{g} = [\mathbf{g}_1^T \ \mathbf{g}_2^T \cdots \mathbf{g}_N^T]^T$, with $\mathbf{g}_n = [\mathrm{g}_{n,0} \ \mathrm{g}_{n,1} \cdots \mathrm{g}_{n,L_i-1}]^T$.

The vector $\mathbf{v}$ is the target vector, i.e. the Kronecker delta shifted by an appropriate modeling delay ($0 \leq d \leq NL_i$), while $\mathbf{H} = [\mathbf{H}_1\mathbf{H}_2 \cdots \mathbf{H}_N]$ and $\mathbf{H}_n$ is the convolution matrix of the RIR between the source and n-th microphone. When the matrix $\mathbf{H}$ is given or estimated through a system identification algorithm, the inverse filter set can be calculated as

$$\mathbf{g} = \mathbf{H}^{\dagger}\mathbf{v} \tag{13}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudoinverse.

Considering the presence of disturbances, i.e. additive noise or RTFs fluctuations, the cost function (12) is modified as follows [2]:

$$C = \|\mathbf{H}\mathbf{g} - \mathbf{v}\|^2 + \gamma \|\mathbf{g}\|^2, \tag{14}$$

where the regularization parameter $\gamma \geq 0$ is a scalar coefficient representing the weight assigned to the disturbance term.

In [2] a general cost function, embedding noise and fluctuations, is derived together with the inverse filter. However, the inverse filter computation requires a matrix inversion that, in the case of long RIRs, can result in a high computational burden. Instead, an adaptive algorithm [10], based on the steepest-descent technique, has been here adopted to satisfy the real-time constraints:

$$\mathbf{g}_n(q+1) = \mathbf{g}_n(q) + \mu(q)[\mathbf{H}^T(\mathbf{v} - \mathbf{H}\mathbf{g}_n(q)) - \gamma\mathbf{g}_n(q)], \tag{15}$$

where $\mu(q)$ is the step-size and q is the time frame index. The convergence of the algorithm to the optimal solution is guaranteed if the usual conditions for the step-size in terms of autocorrelation matrix $\mathbf{H}^T\mathbf{H}$ eigenvalues hold. However, the achievement of the optimum can be slow if a fixed step-size value is chosen. The algorithm convergence speed can be increased choosing a step-size that minimizes the cost function at the next iteration:

$$\begin{aligned} \mu(q) &= \frac{\mathbf{e}^T(q)\mathbf{e}(q)}{\mathbf{e}^T(q)\left(\mathcal{H}^T\mathcal{H} + \gamma I\right)\mathbf{e}(q)}, \\ \mathbf{e}(q) &= \mathcal{H}^T\left[\mathbf{v} - \mathcal{H}\mathbf{g}_{m^*}(q)\right] - \gamma\mathbf{g}_{m^*}(q). \end{aligned} \tag{16}$$

The illustrated algorithm presents two advantages: First, the regularization parameter γ makes the dereverberation process more robust to estimation errors due to the BCI algorithm [2]. Second, the complexity of the algorithm is decreased since no matrix inversion is required and operations can be performed in the frequency-domain through FFTs.

3 Automatic Speech Recognition System

The HMM system applied for processing features computed from the dereverberated speech signal was identical to the back-end used in [15]. 39 cepstral mean normalized MFCC features (including deltas and double deltas) are extracted from the speech signal every 10 ms using a window size of 25 ms. Each phoneme

is represented by three emitting states (left-to-right HMMs) with 16 Gaussian mixtures. The initial monophone HMMs were mapped to tied-state cross-word triphone models with shared state transition probabilities. Two Baum-Welch iterations were performed for re-estimation of the triphone models. Finally, the number of mixture components of the triphone models was increased to 16 in four successive rounds of mixture doubling and re-estimation (four iterations in every round). Both, acoustic models and a back-off bigram language model were trained on the non-reverberated version of the Buckeye training set.

4 Experiments

4.1 Data

Experiments have been conducted on the Buckeye corpus of conversational speech [7]. The corpus consists of interviews of forty native American English speakers speaking in conversational style. Signals have been recorded with close-talking microphones in quiet conditions with a sample rate of 16 kHz. The 255 recording sessions, each of which is approximately 10 min long, were subdivided into turns by cutting whenever the subject's speech was interrupted by the interviewer, or once a silence segment of more than 0.5 s length occurred. We used the same speaker independent training and test sets as in [15]. The lengths of the sets are 23.1 h and 2.6 h, respectively, and the vocabulary size is 9.1 k.

4.2 Experimental Setup

The experimental setup consists of a speaker located in the meeting room shown in Fig. 2. Inside, a table is present and an array of three omnidirectional microphones is located on its centre. Two reverberated conditions have been considered: *Stationary*, where the speaker talks at the seat denoted as "START" for the all duration of the utterance, and *non-stationary*, where the speaker talks at seat "START" for the first 60 s, and at seat "END" for the remaining time. The difference between the two conditions is that in the first the impulse response does not change, while in the second it changes instantaneously.

Three reverberation times (T_{60}) have been considered: 240 ms, 360 ms, and 480 ms. The reverberated test sets have been created concatenating the clean utterances in order to obtain segments with a minimum length of 120 s, and convolving them with the appropriate impulse responses. All the impulse responses are 1024 taps long, and have been generated by means of Habets' RIR Generator tool[1].

Experiments have been conducted on a Intel® Core™i7 machine running at 3 GHz with 4 GB of RAM. In this machine, the C++ implementation of dereverberation front-end achieves real-time execution with a real-time factor of 0.04.

[1] http://home.tiscali.nl/ehabets/rirgenerator.html

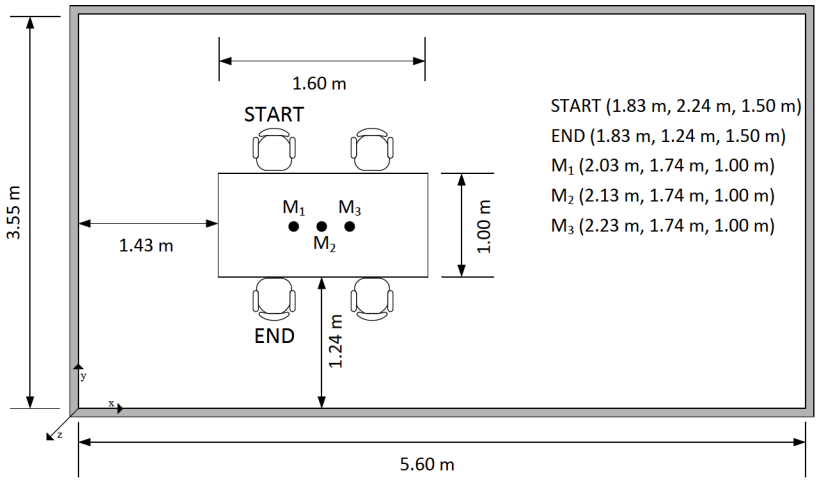


Fig. 2. Room setup: Microphones and speaker's positions coordinates are shown in brackets

4.3 Blind Impulse Response Estimation Performance

The performance of the BCI stage has been evaluated separately to highlight its behaviour in non-stationary conditions. The performance metric used to this end is the NPM [3], defined as:

$$\mathrm{NPM}\,(q) = 20\log_{10}\left(\frac{\|\epsilon(q)\|}{\|\mathbf{h}\|}\right), \tag{17}$$

where

$$\epsilon(q) = \mathbf{h} - \frac{\mathbf{h}^T\widehat{\mathbf{h}}(q)}{\widehat{\mathbf{h}}^T(q)\widehat{\mathbf{h}}(q)}\widehat{\mathbf{h}}(q) \tag{18}$$

is the projection misalignment vector, $\mathbf{h}$ is the real RIR vector whereas $\widehat{\mathbf{h}}(q)$ is the estimated one at the q-th frame.

Fig. 3 shows the NPM curves obtained in the stationary and non-stationary conditions for a Buckeye utterance of length 120 s and reverberated with $T_{60} =$ 480 ms. In stationary conditions, the algorithm reaches an NPM value below -8 dB after about 25 s. In non-stationary conditions, the curve exhibits a peak when the impulse response changes, then starts lowering again reaching a value below -9 dB at the end of the utterance. This shows that the algorithm is able to track the abrupt change of RIRs and it does not suffer from misconvergence. However, a difference of about 2 dB between the stationary and non-stationary NPM curves can be noticed after 30 s from the RIRs change. The behaviour can be explained considering that the BCI algorithm convergence rate depends on the initialization strategy. In this situation, the identification of the "END" impulse response is initialized with the last estimation of the "START" one and not as in equation (10) as indicated in [3].

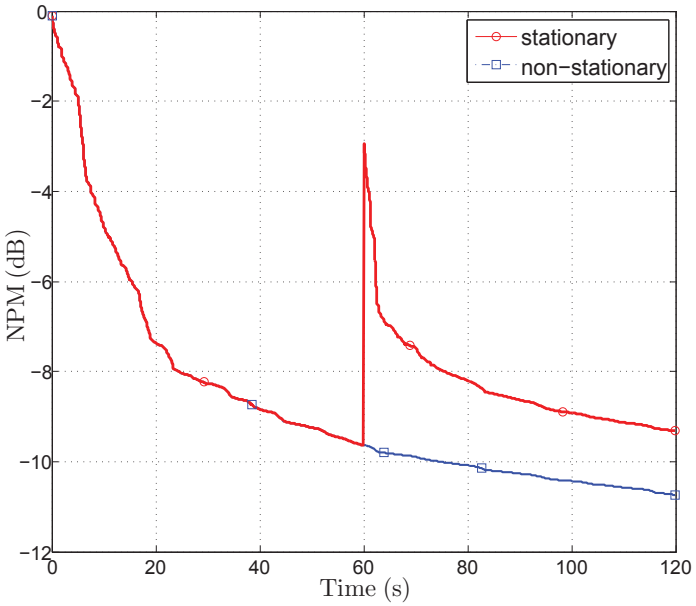


Fig. 3. NPM curves for the stationary and non-stationary conditions

The experiment suggests that a proper re-initialization of the algorithm, e.g. properly dealing with the behaviour of the cost function of equation (14), could be beneficial to the improvement of the tracking capabilities. This idea will be verified in the next section in terms of word recognition accuracy.

4.4 Speech Recognition Results

The word recognition accuracy obtained on clean data without the speech dereverberation front-end is 50.97% (see also [15]). Results obtained on the stationary and non-stationary reverberated conditions are shown in Table 1 for each T_{60}. The column "Average" contains the word accuracy average over the three T_{60}s.

Results show that, without processing, the word accuracy degrades by 12.92% in the stationary case and by 11.91% in the non-stationary one, and that as expected the difference increases with T_{60}. In the stationary case, the dereverberation front-end improves the word accuracy by on average 6.65%, i.e. 17.48% relative. It is worth highlighting that the differences across the three T_{60} are less pronounced: The motivation is that the dereveberation process strictly depends on the quality of the RIRs estimates, thus when a good match between them and the real filter is obtained, the equalization process is effective regardless the reverberation time. In the non-stationary case, the change tracking capability showed in the previous section in terms of NPM are confirmed: The average word accuracy degrades only by 1.28% w.r.t. the dereverberated stationary case, giving a relative improvement of 11.16%.

Table 1. Word accuracy (%) for the addressed conditions

		240 ms	360 ms	480 ms	Average
no processing	stationary	42.71	37.06	34.38	38.05
	non-stationary	43.83	38.30	35.04	39.06
dereverberated	stationary	46.36	44.79	42.95	44.70
	non-stationary	44.84	43.47	41.94	43.42

Re-initializing the BCI algorithm to equation (10) as suggested previously results in a average word recognition accuracy of 44.91%. The re-initialization is performed in an oracle style after the first 60 s of speech, i.e. when the impulse response changes. The word accuracy improves by 1.49% on average w.r.t. the non re-initialized solution and is similar to the dereveberated stationary result. This demonstrates that a proper re-initialization strategy of the BCI algorithm indeed improves the overall performance.

5 Conclusions

In this paper, a speech recognition system able to operate in non-stationary reverberated environments has been presented. The system is composed of a dereverberation front-end and a speech recognition engine able to recognize spontaneous speech. The performance of the front-end has been evaluated in terms of Normalized Projection Misalignment: Results showed that the blind channel identification stage is able to track an abrupt change of RIRs and it does not suffer from misconvergence. The entire system has been evaluated using the Buckeye corpus of conversational speech in stationary and non-stationary environments. In the stationary case, the front-end provides a 17.48% relative word accuracy improvement and the performance is less dependent on the value of T_{60}. In the non-stationary case, the RIRs tracking capabilities are confirmed: The average word accuracy degrades only of 1.28% w.r.t. the stationary scenario. Re-initializing the channel identification algorithm in an oracle style resulted in a average word accuracy improvement of 1.49% demonstrating the effectiveness of the idea.

In future works, the idea of re-initializing the blind channel identification algorithm will be exploited by suitably managing the cost function when the impulse response changes. In addition, the entire system performance will be assessed in different non-stationary conditions, e.g. in a moving-talker scenario. Noise will be addressed modifying the channel identification algorithm [1] and introducing suitable techniques in the speech recognizer feature extraction stage [9,12]. Finally the proposed front-end will be applied in other relevant human-computer interaction scenarios, such as keyword spotting [8,16] and emotion recognition [11].

References

1. Haque, M., Hasan, M.: Noise robust multichannel frequency-domain LMS algorithms for blind channel identification. IEEE Signal Process. Lett. 15, 305–308 (2008)
2. Hikichi, T., Delcroix, M., Miyoshi, M.: Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. EURASIP Journal on Advances in Signal Process. 2007(1) (2007)
3. Huang, Y., Benesty, J.: A class of frequency domain adaptive approaches to blind multichannel identification. IEEE Trans. Speech Audio Process. 51(1), 11–24 (2003)
4. Kumar, K., Singh, R., Raj, B., Stern, R.: Gammatone sub-band magnitude-domain dereverberation for ASR. In: Proc. of ICASSP, pp. 4604–4607 (May 2011)
5. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Signal Process. 36(2), 145–152 (1988)
6. Naylor, P., Gaubitch, N.: Speech Dereverberation. Signals and Communication Technology. Springer (2010)
7. Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E.: Buckeye corpus of conversational speech, 2nd release (2007), http://www.buckeyecorpus.osu.edu, Columbus, OH: Department of Psychology, Ohio State University (Distributor)
8. Principi, E., Cifani, S., Rocchi, C., Squartini, S., Piazza, F.: Keyword spotting based system for conversation fostering in tabletop scenarios: Preliminary evaluation. In: Proc. of 2nd Int. Conf. on Human System Interaction, Catania, pp. 216–219 (2009)
9. Principi, E., Cifani, S., Rotili, R., Squartini, S., Piazza, F.: Comparative evaluation of single-channel MMSE-based noise reduction schemes for speech recognition. Journal of Electrical and Computer Engineering 2010, 6 (2010)
10. Rotili, R., Cifani, S., Principi, E., Squartini, S., Piazza, F.: A robust iterative inverse filtering approach for speech dereverberation in presence of disturbances. In: Proc. of IEEE APCCAS, pp. 434–437 (December 2008)
11. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Communication, 1062–1087 (February 2011)
12. Schuller, B., Wöllmer, M., Moosmayr, T., Rigoll, G.: Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement. EURASIP Journal on Audio, Speech, and Music Processing 2009, 17 (2009)
13. Sehr, A., Maas, R., Kellermann, W.: Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. IEEE Trans. on Audio, Speech, and Lang. Process. 18(7), 1676–1691 (2010)
14. Wölfel, M., McDonough, J.: Distant Speech Recognition, 1st edn. Wiley, New York (2009)
15. Wöllmer, M., Schuller, B., Rigoll, G.: A novel Bottleneck-BLSTM front-end for feature-level context modeling in conversational speech recognition. In: Proc. of ASRU, Waikoloa, Big Island, Hawaii, pp. 36–41 (December 2011)
16. Wöllmer, M., Marchi, E., Squartini, S., Schuller, B.: Multi-stream LSTM-HMM decoding and histogram equalization for noise robust keyword spotting. Cognitive Neurodynamics 5(3), 253–264 (2011)
17. Young, S., Everman, G., Kershaw, D., Moore, G., Odell, J.: The HTK Book. Cambridge University Engineering (2006)

From Nonverbal Cues to Perception: Personality and Social Attractiveness

Alessandro Vinciarelli[1,2], Hugues Salamin[1], Anna Polychroniou[1], Gelareh Mohammadi[2,3], and Antonio Origlia[4]

[1] University of Glasgow, Glasgow, UK
{vincia,hsalamin,annap}@dcs.gla.ac.uk
[2] Idiap Research Institute, Martigny, Switzerland
[3] EPFL, Lausanne, Switzerland
gmohamma@idiap.ch
[4] University of Naples "Federico II", Naples, Italy
antonio.origlia@unina.it

Abstract. Nonverbal behaviour influences to a significant extent our perception of others, especially during the earliest stages of an interaction. This article considers the phenomenon in two zero acquaintance scenarios: the first is the attribution of personality traits to speakers we listen to for the first time, the second is the social attractiveness of unacquainted people with whom we talk on the phone. In both cases, several nonverbal cues, both measurable and machine detectable, appear to be significantly correlated with quantitative assessments of personality traits and social attractiveness. This provides a promising basis for the development of computing approaches capable of predicting how people are perceived by others in social terms.

Keywords: Social Signal Processing, Nonverbal Behaviour, Personality, Social Attractiveness, Voice Quality, Laughter, Back-Channel.

1 Introduction

Social Cognition has shown that unconscious, automatic cognitive processes of which we are unaware have significant influence on our behaviour and attitudes towards others, especially in zero acquaintance scenarios and early stages of an interaction [25,26]. The key aspect of the phenomenon is the *perception-action link* [5], namely the automatic and unmediated activation of behavioural patterns after the very simple perception of appropriate stimuli, whether these correspond to verbal messages (e.g., emotionally or idelogically oriented messages), context and environment characteristics (e.g., wheather and time of the day), or nonverbal behavioural cues (e.g., facial expressions and speaking style) [1,2].

From a computing point of view, the perception-action link is interesting for two main reasons: the first is that it makes human behaviour potentially easier to predict. In fact, if a certain stimulus tends to elicit always the same behavioural pattern, the uncertainty about behaviour under observation can be reduced. For

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 60–72, 2012.

example, when a dyadic interaction participant displays certain nonverbal cues, it becomes easier to predict whether the other participant will back-channel or not [14]. The second reason is that human behaviour can possibly be changed by generating or displaying appropriate stimuli. For example, people that hold certain personality traits tend to spend more time with robots that simulate those same traits [24].

In both cases, the key issue is to understand how people perceive a given stimulus, i.e. what is the social meaning that people tend to attach to it [29]. The reason is that such a meaning seems to determine the behavioural patterns that the stimulus activates [1,2]. In other words, once we have attached a certain meaning to a stimulus, we tend to automatically react to it always in the same way. Hence, this paper considers the way nonverbal behavioural cues typically used in conversations (speaking style, voice quality, prosody, laughter, back-channel, etc.) influence social perception in zero-acquaintance scenarios. In particular, the paper considers two problems: the first is how nonverbal vocal behaviour influences the perception of personality traits in people that listen to a speaker for the first time [6,19,20,22]. The second is the role of laughter, back-channel and turn-taking in shaping the perception of social and task attractiveness when people talk together for the first time [11,12].

According to the Social Signal Processing paradigm [27,28], the ultimate goal of this investigation is the development of approaches capable of predicting automatically not only how individuals perceive one another, but also how they react to the nonverbal cues they mutually display. However, the perception of both personality and social attractiveness has been investigated extensively in psychology as well and the results of this work can provide indications on the role of nonverbal communication in social interactions.

The rest of the paper is organized as follows: Section 2 shows the results obtained in personality perception experiments, Section 3 shows how several nonverbal cues (laughter, back-channel, etc.) shape social and task attractiveness, and Section 4 draws some conclusions.

2 Personality Perception: From Speech to Traits

Personality is the latent construct that accounts for "*individuals' characteristic patterns of thought, emotion, and behaviour together with the psychological mechanisms - hidden or not - behind those patterns*" [6]. Whenever we enter in contact with another person, we quickly develop an impression about her that leads to the attribution of personality traits that, while not being necessarily accurate, still guide our social behaviour, especially in the earliest stages of an interaction [25,26]. This section investigates the effect of nonverbal vocal behaviour (in particular prosody and voice quality) on such phenomenon.

2.1 Measuring Personality: The Big-Five Model

The personality model most commonly applied in the literature, known as the *Big-Five* Model (BF), relies on five broad dimensions that not only capture most

Table 1. BFI-10 questionnaire. The table reports the questions of the BFI-10 and the respective IDs.

ID	Statement	ID	Statement
1	This person is reserved	6	This person is outgoing, sociable
2	This person is generally trusting	7	This person tends to find fault with others
3	This person tends to be lazy	8	This person does a thorough job
4	This person is relaxed, handles stress well	9	This person gets nervous easily
5	This person has a few artistic interests	10	This person has an active imagination

of the observable differences between people, but also are stable across cultures and situations [18]:

- *Extraversion*: Active, Assertive, Energetic, Outgoing, Talkative, etc.
- *Agreeableness*: Appreciative, Kind, Generous, Forgiving, Sympathetic, etc.
- *Conscientiousness*: Efficient, Organized, Planful, Reliable, Responsible, etc.
- *Neuroticism*: Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying, etc.
- *Openness*: Artistic, Curious, Imaginative, Insightful, Original, etc.

Following the lexical hypothesis, adjectives like those in the list above are the physical trace that personality leaves in language [18]. Hence, personality is *measured* by assigning an individual five numerical scores (one per dimension) that account for how well such adjectives describe the person.

The attribution of the scores is typically performed with questionnaires that consider observable behaviour and characteristics of an individual. In this work, we adopted a short version of the *Big-Five Inventory* (see Table 1) [16]. In particular, the assessors involved in the experiments have been asked to answer the questions of Table 1 after listening for 10 seconds to a person they have never heard before (see below for more details). The answers are selected out of a Likert scale with five possible values, from "*Strongly Disagree*" to "*Strongly Agree*", mapped into the inverval $[-2, 2]$. If Q_i is the answer to question i, the scores for the different dimensions are calculated as follows: Extraversion: $Q_6 - Q_1$, Agreeableness: $Q_2 - Q_7$, Conscientiousness: $Q_8 - Q_3$, Neuroticism: $Q_9 - Q_4$, Openness: $Q_{10} - Q_5$. The resuting range for each dimension is $[-4, 4]$.

2.2 Speech and Personality

The interplay between personality and speech takes two main forms. On one hand, our personality is likely to influence the way we speak and leave *markers* in it [20]. On the other hand, our voice quality and speaking style elicit the attribution of certain personality traits rather than others [19]. While the above has been proposed as a hypothesis roughly one century ago [17], quantitative studies have been performed only since the late seventies. The speech features

used in this work are influenced by the results obtained since in the psychological literature and address nonverbal vocal behavioural cues most likely to influence personality perception [22].

The first step of the feature extraction process is the segmentation into *syllables* and *syllable nuclei*. The reason is that these units are less affected by noise and result into more reliable information when processed. In absence of an explicit syllable segmentation, we applied an automatic approach (see [15] for more details) based on the following definition [4, p.275]: "*[a syllable is] a continuous voiced segment of speech organized around one local loudness peak, and possibly preceeded and/or followed by voiceless segments*". The nucleus of the so obtained syllables is the region where the energy is in within $3dB$ from the loudness peak. The length of each syllable is taken as a feature while the ratio between the number of detected syllables and the total duration of an utterance is taken as a measure of speech rate.

Syllable nuclei have been used to extract voice quality related features. The first is *harmonicity*, a measure of the ratio between the energy in the periodic part of the speech signal and the noise (see [3] for the method applied). The second is the *spectral centroid*, perceptually correlated with voice brightness. The centroid is calculated as the average of the frequencies, weighted by their respective energies. The distribution of the energies is used to compute *spectral skewness* (how much energy is above the spectral centroid) and *spectral kurtosis* (how much the energy distribution is different from a Gaussian).

The spectral slope is a measure of the difference between the amount of energy found in the low frequency area and the high frequency area. Spectral slopeness measures have been shown to be effective in emotion discrimination [23]. In this work, spectral tilt is estimated by considering the Long-Term Average Spectrum (LTAS) and taking the slopeness of the trend line computed over the frequency bins. In the employed feature extraction algorithm, the width of the frequency bins is set to 1000 Hz, the low frequency area is comprised between 0 and 1000 Hz and the high frequency area is comprised between 1000 and 4000 Hz. These values are commonly found in the literature.

The frequency values of the first three formants specify the frequencies around which energy concentrates because of the examined vowel. Formant bandwidths describe the area of influence of the considered formants over the spectrum. The two last measurements provide information about syntax-related energy distribution in the syllable nucleus.

Other two voice quality related measures are extracted from syllable nuclei: *Jitter* and *Shimmer*. Jitter is defined as "*the average absolute difference between a period and the average of it and its four closest neighbours divided by the average period*". It is included in the feature set in order to describe the stability of the periodic component inside the syllable nucleus. Shimmer is defined as "*the average absolute difference between the amplitudes of consecutive periods divided by the average amplitude*" and it is included in the features set in order to describe the stability of the energetic component inside the syllable nucleus.

Features concerning the length of syllables and their nuclei are included in the features set to describe the amount of stress in the utterance while energy related features are included as they are a powerful indicator of arousal and dominance levels [21].

The likelihood of dynamic tones (glissando) in a syllable nucleus is estimated for each syllable as the ratio between the actual rate of change of the pitch movement crossing the syllable nucleus and the glissando perception threshold employed in [13]. If the observed rate of change exceeds the threshold, the value of the likelihood is set to 1. This parameter gives an account of whether the pitch movement crossing the syllable nucleus will be perceived as a dynamic tone or as a static one.

For an entire audio clip to be represented, it is necessary to estimate statistical properties of the features above that are extracted from each syllable and nucleus separately. In this work, the statisticals adopted are mean, standard deviation, minimum, maximum and entropy. Different statisticals are used for different features (see caption of Figure 2 for more details).

2.3 Experiments and Results

The goal of the experiments is to show whether the nonverbal cues described in the previous section actually influence the perception of personality or not. A pool of 11 assessors has listened to 640 audio clips of length 10 seconds. The total number of individuals talking in the corpus is 322, with the most represented person speaking in 16 clips and 61% of the individuals speaking only in one clip. For each clip, each of the assessors has filled the questionnaire of Table 1, for a total of 70400 questions answered. The assessments have been performed via an online application and each judge has worked independently of the others. The clips have been presented in a different order to each assessor to avoid tiredness effects. The clips have been assessed in sessions no longer than 30 minutes (no more than two sessions per day) to avoid the lack of concentration resulting from the prolonged repetition of a potentially tedious task. For a given clip, the score for each dimension is the average of the scores assigned by each assessor individually.

The clips have been randomly extracted from the news bulletins broadcasted in Switzerland in February 2005. Only assessors that do not speak the language of the clips (French) have been selected. In this way, the assessors should be influenced only by nonverbal behaviour. The data is emotionally neutral and attention has been paid to avoid words that might be accessible to non-French speakers and have a priming effect (e.g., names of famous people or places) [1,2].

Figure 1 shows the distribution of the scores across the clips of the corpus. For certain traits (in particular Extraversion and Conscientiousness) the assessments cover a large fraction of the range with sufficient frequency. For others, the distribution is peaked around 0, the value corresponding to the answer "*Neither agree nor disagree*". These results are not surprising because Extraversion and Conscientiousness are well known to be perceived quickly and effectively in the

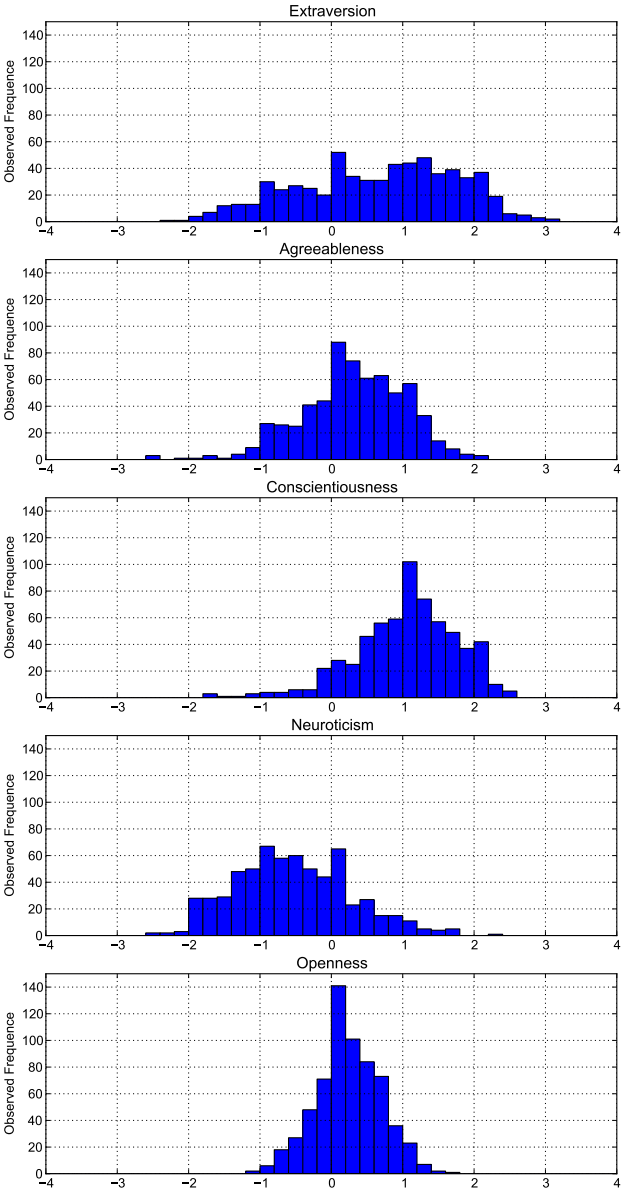


Fig. 1. Distribution of personality scores across different traits

very first instants after a first encounter [8]. In contrast, the other dimensions are difficult to assess in zero acquaintance scenarios.

Figure 2 shows the correlation between speech cues (see previous section) and personality scores. The horizontal lines in the bar charts correspond to a significance level of 1%. For all traits, except Openness, at least half of the

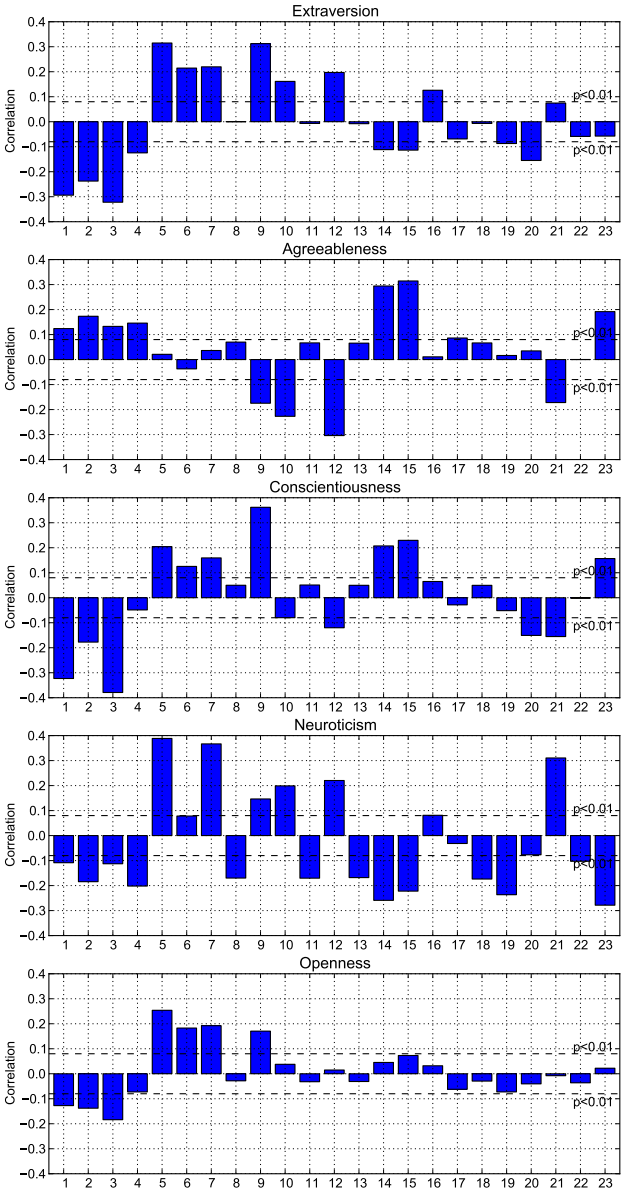


Fig. 2. Correlation between features and personality scores: Nuclei Len. Mean and Entropy (1,2), Syllables Len. Mean and Entropy (3,4), F0 Mean, Stdev., Minimum and Entropy (5,6,7,8), Speech Rate (9), Spectral Centroid and Entropy (10,11), Spectral Tilt and Entropy (12,13), Spectral Skewness and Kurtosis (14,15), Energy Mean, Maximum and Entropy (16,17,18), Jitter and Shimmer (19,20), Mean of F1, F2 and F3 (21,22,23).

cues are correlated with p-value lower than 1%. However, the cues with higher correlation change depending on the trait. The perception of Extraversion seems to be dominated by length of syllables and vowels (the shorter syllables and vowels, the higher the assigned Extraversion score), pitch (the higher the pitch, the higher the perceived Extraversion) and speaking rate (the faster the speaker, the more extrovert she sounds). In contrast with the psychological literature, the energy seems not to have an effect, but this is probably due to the relatively small variability of loudness in the corpus.

In the case of Agreeableness, spectral cues dominate perception: brighter voices (higher center of mass in the power spectrum) and higher spectral tilt (higher energy fraction on the fundamental frequency) are perceived as less agreeable. In contrast, voices for which the power spectrum is peakier and tends to be skewed towards higher frequencies are perceived as more agreeable. These latter cues affect the perception of Conscientiousness in the same way, together with the speaking rate (people that talk faster look more competent). In the case of Neuroticism, the higher pitch and first formant means, the higher the score. No evident effects are observed for Openness and the reason is probably that this trait is difficult to assess in a scenario like the one considered in this work, as it is evident in the score distribution narrowly peaked in correspondence of the "*Neither agree nor disagree*" answer.

3 Nonverbal Communication and Social Attractiveness

No other technologies have been accepted as widely as cellular phones in everyday life. In 2005, a mere 15 years after their first apperance in the consumer electronics market, there was one mobile phone subscription every third person in the world, with 82 subscriptions per 100 persons in Europe (the most "mobile" continent) and 19 countries where the number of subscriptions exceeded the size of the population (see [7,9] for up-to-date figures). The ubiquitous diffusion of mobile phones is a major change in the way we develop and maintain our social ties [10]. However, the impact on conversation, the primary site of human sociality, has not been investigated extensively. The results presented in this section try to address such a gap by showing how a number of nonverbal behavioural cues influence the perception of social and task attractiveness [11,12] between unacquainted people talking on the phone.

3.1 Data and Scenario

The experiments of this section have been performed over a collection of 26 phone calls between unacquainted individuals. The total number of involved subjects is 52 (no person participates in more than one call). During the data collection, the subjects are invited to the laboratory, but they do not meet one another before the call. The conversations are centered around the *Winter Survival Scenario* (WSS): the two persons play the role of members of a rescue team that must support a group of people that have survived a plane crash in Northern Canada.

Table 2. Social (left column) and task (right column) attractiveness questionnaires

ID	Statement	ID	Statement
1	I think (s)he could be a friend of mine	1	I couldn't get anything accomplished with him (her)
2	I would like to have a friendly chat with him (her)	2	(S)he is a typical goof off when assigned a job to do
3	It would be difficult to meet and talk with him (her)	3	I have confidence in his (her) ability to get the job done
4	We could never establish a personal firednship with each other	4	If I wanted to get things done, I could probably depend on him (her)
5	(S)he just would not fit into my circle of friends	5	(S)he would be a poor problem solver
6	(S)he would be pleasant to be with	6	I think studying with him (her) would be impossible
7	I feel I know him (her) personally	7	You could count on him (her) getting the job done
8	(S)he is personally offensive to me	8	I have the feeling (s)he is a very slow worker
9	I do not care if I ever get to meet him (her)	9	If we put our heads together, i think we could come up with some good ideas
10	I sometimes wish I were more like him (her)	10	(S)he would be fun to work with

It is winter (temperatures around -40^oC) and the survivors have extracted 12 items from the plane. However, they have to leave the place of the crash and they can bring only part of the 12 items. During the call, the rescue members are expected to identify the items that maximize the chances of survival. The subjects are paid 6 British Pounds for their participation. Furthermore, they get 3 extra Pounds each time they select a good item, but are penalized by the same amount each time they select a wrong one (in any case, a minimum payment of 6 Pounds is guaranteed).

The conversations have been captured with two cellular phones (Nokia N900) that record not only what the subjects say (via both microphone and speakers), but also the movement of the phones via accelerometers, gyroscopes and magnetoscopes. All signals are synchronized to allow a multimodal analysis of subjects behaviour. The conversations have been annotated manually in terms of *turns* (who speaks when), *silences* (when none of the subjects talks), *laughter* (both individual and common) and *back-channel* (short vocal bursts that react and/or accompany the speech of others without attempts of grabbing the floor).

Furthermore, the subjects have filled the questionnaires proposed in [11,12], aimed at assessing both social (how much we enjoy interacting with another person) and task (how much we like to work with another person) attractiveness of their interlocutor (see Table 2). Like in the case of the BFI-10, the questions are associated to Likert scales with five possibile answers, from "*Strongly Disagree*"

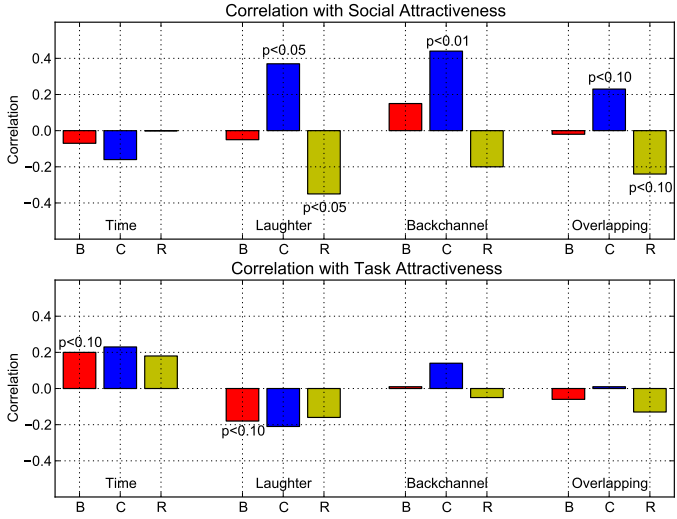


Fig. 3. Recognition performance as a function of the clips length. The right plot shows the results for the two classes separately.

to "*Strongly Agree*", mapped into the interval $[-2, 2]$ (see Table 2 for more details). If Q_i is the answer to question i, the social attractiveness is calculated as $Q_1+Q_2-Q_3-Q_4-Q_5+Q_6+Q_7-Q_8-Q_9+Q_{10}$. In the case of the task attractiveness, the sum is as follows: $-Q_1-Q_2+Q_3+Q_4-Q_5-Q_6+Q_7-Q_8+Q_9+Q_{10}$.

3.2 Experiments and Results

Figure 3 shows the correlation between the nonverbal cues annotated in the data and the attractiveness scores obtained from the questionnaires. The cues considered are the fraction of time a person talks during a call, the number of times a person laughs, the number of times a person performs back-channel, and the number of times there is overlapping speech. For each cue, the correlation is calculated for all speakers (B), only for the subjects that call (C) and only for the subjects that receive (R).

The upper chart shows the results for the social attractiveness. When all speakers are taken into account, no correlation reaches a p-value lower than 10%. However, the situation changes when considering separately the subjects that call and those that receive. The former tend to be more socially attractive when they laugh more ($p < 5\%$), when they show more back-channel ($p < 1\%$) and when there is more overlapping speech ($p < 10\%$). In the case of the subjects that receive the situation is opposite, namely they are more appreciated if they laugh less ($p < 5\%$) and there is less overlapping speech ($p < 10\%$). With the exception of the fraction of time people talk, the results seem to suggest that the expectations are different depending on whether a person calls or is called.

However, the effect might depend on the particular scenario adopted and it should be confirmed by collecting more data.

In the case of task attractiveness, no correlations reach a p-value lower than 10% and it is not possible to say whether there is a difference between being the person that calls or the one that receives. However, correlations with acceptance level lower than 10% are obtained when considering all of the subjects. In particular, people that talk more, but laugh less, seem to be more attractive when it is necessary to accomplish a task. In this case as well, the number of calls (26) and subjects (52) is relatively low and more solid evidence can be obtained only by collecting more data.

4 Conclusions

This paper has investigated how people perceive a number of nonverbal behavioural cues in social terms. Two problems have been considered: the first is how people attribute personality traits to speakers they have never heard before. The second is the perception of social and task attractiveness in phone calls between unacquainted individuals.

In the first case, the experiments have focused on voice and speaking style characteristics (voice quality, speaking rate, etc.) and show that a large number of nonverbal cues correlate to a statistically significant degree with personality assessments. In the second case, laughter and backchannel have been shown to influence significantly the perception of social attractiveness, but they do it in a different way depending on whether a person calls or is called. To the best of our knowledge, it is the first time that someone reports about such a lack of symmetry between two people involved in a phone conversation.

The results presented in this work are interesting under two main respects. The first is that they provide further information about nonverbal cues involved in important social phenomena. The second is that they provide useful indications about the cues to be detected in order to develop automatic approaches capable of predicting the way people perceive others.

Acknowldgments. The research that has led to this work has been supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet), in part by the Swiss National Science Foundation via the National Centre of Competence in Research IM2 (Information Multimodal Information management) and in part by the Finnish Funding Agency for Technology and Nokia Research Center Finland as part of Human Emotional Interaction project 2010-2011.

References

1. Bargh, J.A., Chen, M., Burrows, L.: Automaticity of Social Behaviour: Direct Effects of Trait Construct and Stereotype Activation on Action. Journal of Personality and Social Psychology 71(2), 230–244 (1996)

2. Bargh, J.A., Williams, E.L.: The Automaticity of Social Life. Current Directions in Psychological Science 15(1), 1–4 (2006)
3. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program]. Version 5.2.40 (2011)
4. D'Alessandro, C., Mertens, P.: Automatic pitch contour stylization using a model of tonal perception. Computer Speech and Language 9(3), 257–288 (1995)
5. Dijksterhuis, A., Bargh, J.A.: The Perception-Behavior Expressway: Automatic Effects of Social Perception on Social Behavior. In: Zanna, M.P. (ed.) Advances in Experimental Social Psychology, pp. 1–40 (2001)
6. Funder, D.C.: Personality. Annual Review of Psychology 52, 197–221 (2001)
7. ITU. The World in 2010: ICT Facts and Figures. Technical report, International Telecommunication Union
8. Judd, C.M., James-Hawkins, L., Yzerbyt, V., Kashima, Y.: Fundamental Dimensions of Social Judgment: Unrdestanding the Relations Bteween Judgments of Competence and Warmth. Journal of Personality and Social Psychology 89(6), 899–913 (2005)
9. Kalba, K.: The Global Adoption and Diffusion of Mobile Phones. Technical Report December, Center for Information Policy Research Harvard University (2008)
10. Ling, R.: New Tech, New Ties. How Mobile Communication is Reshaping Social Cohesion. MIT Press (2008)
11. McCroskey, J.C., McCain, T.A.: The measurement of interpersonal attraction. Speech Monographs 41(3), 261–266 (1974)
12. McCroskey, L., McCroskey, J., Richmond, V.: Analysis and Improvement of the Measurement of Interpersonal Attraction and Homophily. Communication Quarterly 54(1), 1–31 (2006)
13. Mertens, P.: The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In: Proceedings of Speech Prosody (2004)
14. Morency, L.P., de Kok, I., Gratch, J.: A Probabilistic Multimodal Approach for Predicting Listener Backchannels. Journal of Autonomous Agents and Multi-Agent Systems 20(1), 70–84 (2010)
15. Petrillo, M., Cutugno, F.: A syllable segmentation algorithm for english and italian. In: Proc. of Eurospeech, pp. 2913–2916 (2003)
16. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality 41(1), 203–212 (2007)
17. Sapir, E.: Speech as a personality trait. The American Journal of Sociology 32(6), 892–905 (1927)
18. Saucier, G., Goldberg, L.R.: The language of personality: Lexical Perspectives on the Five-Factor Model. In: Wiggins, J.S. (ed.) The Five-Factor Model of Personality (1996)
19. Scherer, K.R.: Personality inference from voice quality: The loud voice of extroversion. European Journal of Social Psychology 8, 467–487 (1978)
20. Scherer, K.R.: Personality markers in speech. In: Social Markers in Speech, pp. 147–209. Cambridge University Press, Cambridge (1979)
21. Scherer, K.R., Johnstone, T., Klasmeyer, G.: Vocal expression of emotions. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (eds.) Handbook of Affective Sciences, pp. 433–456. Oxford University Press (2003)
22. Scherer, K.R., Scherer, U.: Speech behavior and personality. In: Speech Evaluation in Psychiatry, pp. 115–135. Grune & Stratton, New York (1981)

23. Tamarit, L., Goudbeek, M., Scherer, K.: Spectral slope measurements in emotionally expressive speech. In: Proceedings of Speech Analysis and Processing for Knowledge Discovery (2008)
24. Tapus, A., Tapus, C., Matarić, M.: User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. Intelligent Service Robotics 1(2), 169–183 (2008)
25. Uleman, J.S., Newman, L.S., Moskowitz, G.B.: People as flexible interpreters: Evidence and issues from spontaneous trait inference, vol. 28, pp. 211–279. Elsevier (1996)
26. Uleman, J.S., Saribay, S.A., Gonzalez, C.M.: Spontaneous inferences, implicit impressions, and implicit theories. Annual Reviews of Psychology 59, 329–360 (2008)
27. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an emerging domain. Image and Vision Computing Journal 27(12), 1743–1759 (2009)
28. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schroeder, M.: Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. IEEE Transactions on Affective Computing (to appear, 2012)
29. Wharton, T.: Pragmatics and Non-Verbal Communication. Cambridge University Press (2009)

Measuring Synchrony in Dialog Transcripts

Carl Vogel and Lydia Behan

Computational Linguistics Group,
School of Computer Science and Statistics,
Trinity College, Dublin 2, Ireland
{vogel,behanl}@tcd.ie

Abstract. A finite register method of processing dialog transcripts is used to measure interlocutor synchrony. Successive contributions by participants are measured for word n-gram repetitions and temporal overlaps. The Zipfian distribution of words in language use leads to a natural expectation that random re-orderings of dialog contributions will unavoidably exhibit repetition – one might reasonably expect that the frequency of repetition in actual dialog is in fact best explained as a random effect. Accordingly, significance is assessed with respect to randomized contrast values. The contrasts are obtained from averages over randomized reorderings of dialog contributions with temporal spans of the revised dialogs guided by the original durations. Benchmark distributions for allo-repetition and self-repetition are established from existing dialog transcripts covering a pair of pragmatically different circumstances: ATR English language "lingua franca" discussions, Air-Traffic communications (Flight 1549 over the Hudson River). Repetition in actual dialog exceeds the frequency one might expect from a random process. Perhaps surprisingly from the perspective of using repetition as an index of synchrony, self-repetition significantly exceeds allo-repetition.

1 Background

Research into synchrony in dialog has deployed methods such as introducing delay, by using video to mediate communication, in a way that allows manipulation of whether interlocutors have access to partner contributions in real-time or with constructed delay. One type of study involves mother-infant communications mediated by video. The experimental paradigm makes the delay absolute, with a re-play condition seamlessly edited in between live interaction phases [15]. The striking effect is the disinterest expressed by the infant when the delay costs the illusion of interaction.

We focus here on two potential factors in the perception of interaction. It would be unsurprising if a politician revealed tactics for seeming engaged during meetings with the public, even when thinking about other matters entirely, as including occasionally repeating words or phrases uttered by their interlocutor or timing contributions to occasionally seem so interested in the content of the conversation to intervene through interruption or talk at the same time as interlocutors without actually taking the floor. In fact, repetition as an indication of

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 73–88, 2012.

participatory listening is just one of many functions of repetition in conversation [17]. Our inspection of interaction focuses on these measures. If these tactics are successful, then it must be because measures of lexical and sub-lexical repetition and of temporal overlap in natural conversations are significantly different than in un-natural conversation. Obviously, a conversation in which two interlocutors talk to each other about different topics would be quite un-natural, and any lexical overlap would be a matter of chance. Nonetheless, in natural dialog, it may be the case that overlaps and lexical repetitions are randomly distributed.

Even in reflecting on the fact that language is not a random process, considering that word rank-frequency distributions are Zipfian, one might reach the conclusion that repetitions are inevitable, and that the chances of repetition between dialog contributions are, in fact, best described by coin tosses. Such a position is motivated perhaps by the conviction that even if interlocutors are cooperatively talking about the same issue, they do so as independent agents, each making their own lexical decisions. Recent work has attempted to demonstrate the extent to which repetition in dialog correlates with task success [12,13].

To explore these questions with a Monte Carlo approach,[1] from natural dialog transcripts we construct randomized versions of the turns. In what follow, "a turn", "a contribution" and "a line" are synonymous expressions. They make sense when thinking of dialog transcriptions as scripts (within which actors have lines). Randomization of turns refers to re-ordering the turns with respect to each other, temporally, but not reordering within any turn. In the experiments reported here, we randomize real dialogs ten times. This means that the contributions of each participant are parsed into a data-structure in sequence from an actual transcript, and then each turn in the actual sequence is assigned a time-stamp in a range determined by the actual overall conversation duration. In the re-ordered dialogs, speakers still say the same thing overall, but not with any semblance of actual synchronization of contributions with respect to each other. Using this experimental framework of comparing measures in actual dialogs with counterpart measures averaged over randomized re-orderings, we focus on measures of lexical repetition by the speaker of others' most recent contributions (allo-repetition), their own most recent contributions (self-repetition), and temporal overlap of contributions. With the focus in this methodology on repetition of components of most recent contributions, a fixed period is searched for potential repeated content. However, the actual temporal durations involved between a given contribution and the contributions which immediately precede it by each speaker may vary. The method we use contrasts with more powerful recurrence analysis techniques for identifying temporal coupling [14]; here, the window for anticipated repetition is structurally rather than durationally restricted.

In what follows, we first describe our register-based method of analyzing individual dialog contributions and durations with respect to the immediately prior turn of each speaker. We also describe the method of constructing randomized re-orderings of the turns by re-assigning pseudo-randomly determined

[1] The method fits into the "full-sample" dimension of the categorization provided by [11] with the labels "randomization", "permutation" and "shuffling".

start and stop times for each contribution. Within exemplars of several types of conversations, we assess levels of self-repetition and allo-repetition, towards an understanding of what counts as unmarked levels for both measures. We provide two case studies of analysis in this paradigm: §3.1 analyzes transcripts of dialogs involving five people over three sessions with a relevant feature that English provides a common ground; §3.4 examines a transcript of air-traffic communications, because extensive repetition is expected in this sort of dialog. We analyze the data with respect to levels of allo-repetition and self-repetition, and where possible, with respect to temporal overlap.

We find strong effects that separate actual dialog from randomized dialog: lexical expressions (from single words to sequences of up to five elements) are more likely to be repeated between contributions in actual dialog than in randomized dialog. Moreover, self-repetition effects are even stronger than repetition of others in ordinary chat. Overlap with others is also distinctive in actual dialog.

2 Methods

We analyze dialogs that have already been transcribed and are available on the web. Therefore, one issue of treatment that we do not have to address in this paper is the tokenization of the recorded dialog into tokens that might be deemed individual contributions of the speakers (1). At some point, one must make a decision between one contribution of Speaker $\mathcal{A}$ that has another speaker overlapping in the middle, and two separate contributions by Speaker $\mathcal{A}$. Making these decisions about the units of dialog constituents contributed by each speaker cannot be easy, and we do not revise any of the transcribers' decisions in this regard. We take a "line" of dialog to be an individual contribution of a speaker as attributed by a transcriptionist, the "lines" of a dialog is the partially ordered sequence of interleaved contributions. The transcripts are temporally ordered, but not totally so, given that contributions of interlocutors are interleaved. Each file is processed using a 'register' (3) for each speaker, initially empty, containing the contents of their most recent contribution (2).

(1) Ξ is the cast of actors (α) communicating.
(2) $u_j = \langle \tau_b, \tau_e, \alpha, \sigma \rangle$ is the j-th transcribed utterance:
τ_s, start time; τ_e, end time; α, actor; σ, statement
At u_j, $\alpha^{u_j} = actor(u_j)$; $\sigma^{u_j} = statement(u_j)$; etc.
(3) $\mathcal{R}_\alpha$, for each $\alpha \in \Xi$, is a register that records the start time, stop time and content of the last utterance of α.
$\mathcal{R}_\Xi$ refers to the set of registers;
$\mathcal{R}_{\Xi/\alpha}$ refers to the set of registers for all actors but α;
$(g(n,\sigma)[i]$ denotes the i-th element of $g(n,\sigma))$.

Ultimately, for each utterance, count tokens shared with immediately preceding turns (their own (6), and their interlocutors' (5), in both cases, as given by (4) for the each length of $n-$gram to be counted) as recorded in the interlocutors' registers. The actual repetition values are then compared with those derived from

some number (ten in the experiments reported here) of randomized re-orderings of the turns (AKA contributions). The constituent sequence of words within any individual contribution are left intact in their original order.

(4) $\kappa(n, \sigma^1, \sigma^2) = \sum_{i=1}^{g(n,\sigma^1)[max]}(g(n, \sigma^1)[i] \in g(n, \sigma^2))$
(5)

$$\textit{allo-shared}(u_j, \mathcal{R}_\Xi, n) = \sum_{\alpha}^{\Xi/\alpha^{u_j}} \kappa(n, \sigma^{u_j}, \sigma^{\mathcal{R}_\alpha})$$

(6) a. first count with respect to former value for self:

$$\textit{self-shared}(u_j, \mathcal{R}_\Xi, n) = \kappa(n, \sigma^{u_j}, \sigma^{\mathcal{R}_{\alpha^{u_j}}})$$

b. update register for the agent's own current utterance:

$$\mathcal{R}_{\alpha^{u_j}} := \langle \tau_s^{u_j}, \tau_e^{u_j}, \sigma^{u_j} \rangle$$

The random-reordering of the dialogs is effected by generating new start-times and durations for each utterance, and then sorting the utterances on their temporal indices. The times are selected using using random generators based on parameters that depend on the values in the original conversation (7). Thus, for each utterance u_i a re-indexing u'_i is constructed (8).

(7) given $u_1 \ldots u_{max}$
   a. $\textit{starttime} = \tau_s^{u_1}$
   b. $\textit{stoptime} = \tau_e^{u_{max}}$
   c. $\textit{maxoverlap}$ = maximum temporal overlap of u_i and $\mathcal{R}_\Xi$ noted at time of shared n-gram computation.

(8) for each u_i,
   a. $\tau_s = \textit{rand}(0, \textit{stoptime})$
   b. $\tau_e = \tau_s + \textit{rand}(0, \textit{maxoverlap})$
   c. $u'_i = \langle \tau_s, \tau_e, \alpha^{u_i}, \sigma^{u_i} \rangle$

The u' are sorted on their value for τ_s. In the re-ordered dialog, we measure overlap, allo-shared tokens and self-shared tokens as before. Again, we consider n-gram sequences up to $n = 5$, and contrast "reality" with 10 randomizations.

Thus, an output file is generated which contains a temporal sequence of lines, each annotated for exactly one speaker and with appropriate measures for each line. One such line is constructed for each level of N-grams up to five. Repetitions of N-grams are recorded as counts with respect to the values in the registers as either SELFSHARED or OTHERSHARED tokens. Here we focus on these as count values as opposed to ratios that relativize figures to the total number of N-grams that would have been possible to share between a dialog contribution and preceding contributions stored in registers.[2] Durations in seconds of temporal

[2] Given that we consider tokenization up to $N = 5$, we wanted to simplify treatment of utterances with fewer than N tokens, and thus avoid domain errors from division by zero.

overlaps with the most recent contributions of each speaker are also recorded for each line, as well as the count of the number of overlaps given the timestamps in each of the registers at the point of evaluation.

The annotations of times on the randomized dialogs interact with the computation of potential overlap in the DialogType=RANDOMIZED conditions. A start-time for each utterance is selected between 0 and the total number of seconds in the day's actual conversations. A corresponding stop-time for each utterance is recorded as the utterance's start-time plus an offset given by a random number of seconds between 0 and 10, the maximum overlap duration within the first dialog's actual conversations. Each shuffled dialog is determined by the old contributions sequenced by their new temporal order.

3 Application of Methods

3.1 Case Study 1: English as *Lingua Franca* in Balanced Chat

The data used here is that described by Campbell [5], conversations in English over three days inclusive of 5 speakers (two native-English speaking), at ATR in Japan.[3] The conversations are relatively balanced in the contributions made by participants, and when watches the accompanying video, one notices a high level of mutual engagement. It is reasonable to take these conversations as representative of engaged, balanced conversation. One speaker, g, was present only for the second day. We treated the data solely by regularizing the time-stamps to an HH:MM:SS format. We extracted columns of the data corresponding to timestamp, speaker id, and transcribed speech. We did not alter the transcriptions: editing errors are not addressed, nor are records of transcription difficulty ("@w") updated. The former would regularize spelling, and would increase the chances of repetition in all conditions. The latter would diminish current instances of repetition, separating the individual instances of the marker into more clear words. However, an interesting feature of the current edition of the transcripts is that the marker encodes unintelligibility of contributions. Whether speaker decision about utterance effort resulting in unintelligibility is conscious in these instances is a matter for debate; however, general effects of speaker intelligibility during the course of dialog are known (eg. Bard et al. [2]). There are 29900 lines of real dialog, and another 299000 from randomizations.

3.2 Results

Shared Expressions. A binary variable, DialogType, records whether the measurements for an item correspond to a dialog contribution in its actual order or in a random one. In analyzing the data, the four levels of N greater than one were coalesced into a single level ("2+") of a related variable N'. Using a generalized linear model with a quasi-poisson error distribution, we separately

[3] We are grateful to Nick Campbell for use of the data from `http://www.speech-data.jp` – last verified February 2012.

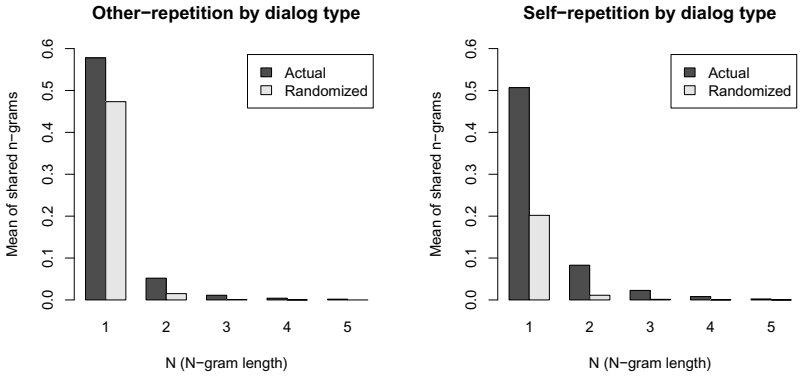


Fig. 1. *Actual* dialog vs *Randomized* turns: sharing others' (left) & own N-grams (right), by N

considered all interactions of DialogType × DAY × Speaker × N' on OtherShared and then on SelfShared. Figure 1 (L) graphs the distribution of mean scores for the count of OtherShared, and Fig. 1 (R) shows the same for SelfShared, both for each value of N. Significantly higher values for each value of N obtain in the actual dialogs than in the randomized dialogs for both repetition of others and of self. The effect of DialogType being set to actual in contrast to the randomized contrast is significantly higher values of OtherShared ($p < 0.005$) and of SelfShared ($p < 2 * 10^{-16}$). Interactions that do not include the factor DialogType=actual are not of interest: effects that obtain or which are commented upon as not emerging include the interaction with DialogType=actual. With respect to repetition of sequences in the preceding contributions of the others, the four-way interactions are not significant, nor the three way interactions. The actual orderings combined with $N' = 2+$ have the effect of significantly higher counts of OtherShared ($p < 3.1 * 10^{-5}$). No effects of DAY or Speaker emerged. Considering self-repetitions, there were significant positive effects of Speaker for g ($p < 0.02$) and $N' = 2+$ ($p < 1.3 * 10^{-9}$). There was a positive interaction for Speaker g with $N' = 2+$ ($p < 0.009$), and negative interaction for Speakers k and y with $N' = 2+$ ($p < 0.05$).

Figure 2 (L) shows the means of repetition of others by speakers, and (on the right) the means of self-repetitions. Self-repetition is systematically greater than repetition of others in the difference from the random values. Figure 3 shows continuity of the main effects over the three days. Interactions of effects are shown in Figures 4 and 5. In actual dialogs, the mean of repetitions for each day, speaker or choice of N in N-gram counts is at least the level in the randomized dialogs, or at a higher level. This holds for both allo-repetition and self-repetition, but the self-repetition values yield greater differences to the randomized values.

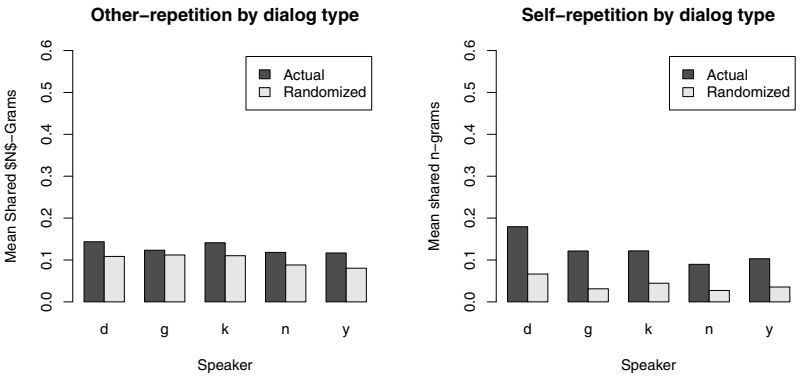


Fig. 2. Repetition in *Actual* vs *Randomized* dialog, of: others (L) & self (R), by Speaker

Temporal Overlap. The approach to simulated overlap adopted in here is naive in allowing the contribution durations in seconds to be a random number between 0 and the longest overlap time. While this might seem to install a proclivity away from overlaps, it actually does not, as is illustrated in Figure 6. In any case, the distribution of actual overlaps is more sharply skewed than the overlaps in the randomized data. The actual dialogs show significantly less overlap than the randomized ($p < 2 * 10^{-16}$) dialogs. SPEAKER n exhibits significantly less overlap $p < 0.001$ than the randomized controls.[4] (The lesser amount of overlap involving speaker k approached significance.) Increased overlap was exhibited on Day 2 ($p < 2 * 10^{-16}$) and decreased overlap on Day 3 ($p < 1.5 * 10^{-6}$). An interaction with Day 2 and SPEAKERS k and n involves greater overlap for both ($p < 0.001$), and Day 3 for SPEAKER n ($p < 0.001$). No other factors studied have significant interactions jointly or in isolation with the condition where DIALOGTYPE=ACTUAL. That the actual data diverges so sharply from the random data may be an artifact of the particular simulation strategy used. Obviously, all of the effects reported are artifacts of the simulation strategy; however, alternative methods of assigning random temporality merit exploration.

3.3 Discussion

The direction of difference in actual dialogs between self-repetition and repetition of other interlocutors in the results is perhaps surprising. An analysis that was thus not anticipated at the outset reveals that the difference is significant. While the results reported show that the means allo-repetition and self-repetition are close, the difference between the means for allo-repetition and the randomized counterpart is smaller than the difference between actual self-repetition and its randomized counterpart. To quantify this, we constructed a

[4] `glm(OverlapSeconds~DialogType*Speaker*Day)`, quasipoisson error family.

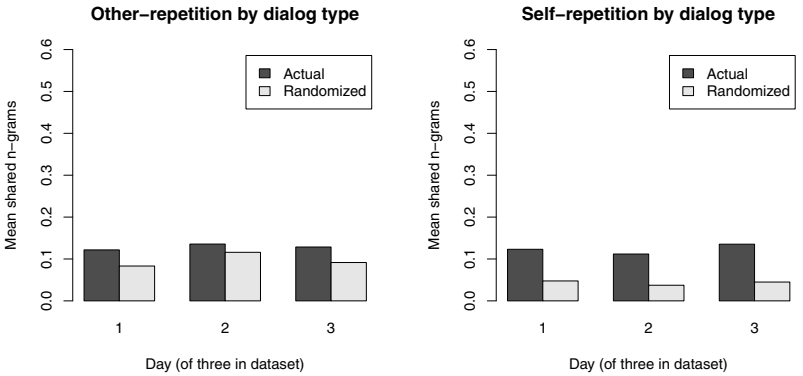


Fig. 3. Repetition in *Actual* vs *Randomized* dialog, of: others (L) & self (R), by Day

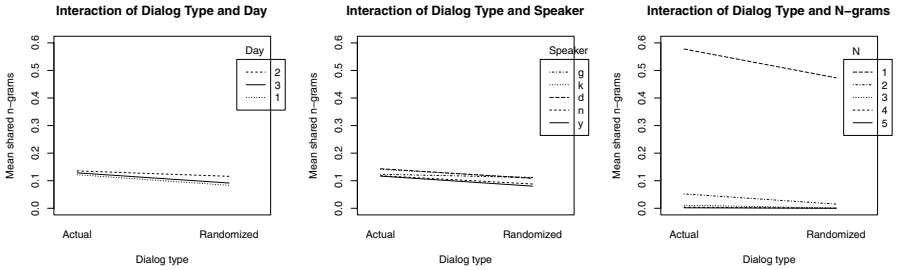


Fig. 4. Interactions of repetitions of others' N-grams: Dialog-Type vs Day (left), Speaker (middle), N (right)

variable DSHARED as SELFSHARED - OTHERSHARED. Our reasoning was that if our perception that self-repetition is stronger than other-repetition, then the effects should be visible in this constructed variable. We reason that if the difference between SELFSHARED and OTHERSHARED in the actual conversation is positive and significantly bigger than the randomized counterpart, then we have captured a difference that separates SELFSHARED from its randomized version as greater than OTHERSHARED and its random counterpart. If the DSHARED value is negative, and the difference between the real and random version is significant, then it is the allo-repetition value that provides the greater difference (and significantly so). Our proxy measure of the relationship we are actually interested in is not the only possible one available to evaluate. We then tested effects on this variable using a generalized linear model with a Gaussian error family.[5] There is, in fact, a significant positive effect of DIALOGTYPE=ACTUAL on this variable ($p < 2 * 10^{-16}$).

[5] `glm(DShared~DialogType)`.

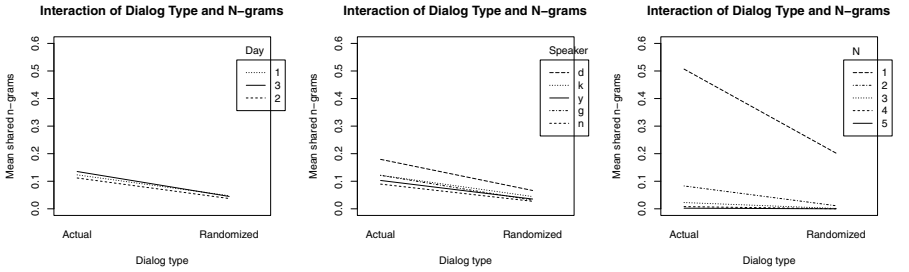


Fig. 5. Interactions of self-repetitions: Reality vs Day (left), Speaker (middle), N (right)

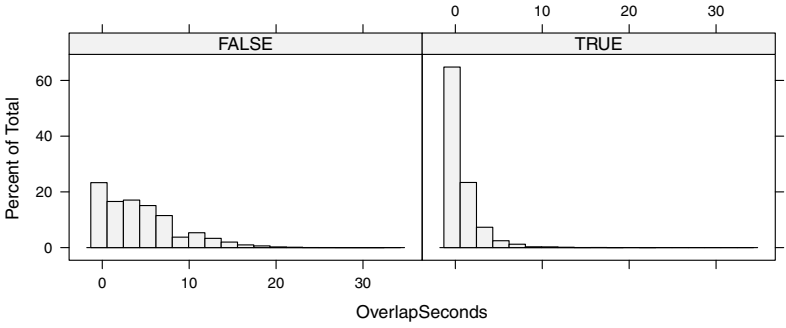


Fig. 6. Dialog Type *Actual* vs. *randomized* Mean Temporal Overlap Histograms

3.4 Case Study 2: A Crisis Situation with a Pre-defined Leader

Given assumptions about repetition in air traffic communications (extensive) and overlaps (scarce), we decided to analyze a dialog from this setting. It happens that transcripts where incidents are involved are most readily available. We decided to examine the one recorded during the landing of US Airways Flight 1549 on the Hudson River on January 15, 2009. The transcript was prepared from the cockpit voice recorder [4]. Following consultation with a licensed flight instructor, we ignored all of the automated contributions; one of those, that of the Automated Terminal Information System (ATIS), is on a loop and the pilot must repeat "papa" when the relevant information is registered. There are 2860 dialog contributions in the resulting corpus. Thus, we consider the contributions of 18 of the recorded channels: CAM, CAM-?, CAM-1, CAM-2, CLC, DEP, GND, HOT-?, HOT-1, HOT-2, INTR-1, INTR-4, PA-1, PA-2, RDO-1, RDO-2, RMP, TWR. Here, we ignore the fact that individuals choose different channels for different purposes (e.g. PA-1 and INTR-1 are both the captain, speaking to the passengers in one case and to ground crew in the other). In the analysis reported below, Voices are therefore individuated as the distinct sources: RDO-1,

CAM-1, PA-1, HOT-1 and INTR-1 all contain the voice of the captain; RDO-2, CAM-2, PA-2, HOT-2 and INTR-2 all contain the voice of the first officer; the other channels are all analyzed as "AllElse".

Results. Fig. 7 shows the mean sharing of N-grams between the ACTUAL and RANDOMIZED dialog types for allo-repetition and self-repetition. It can be seen that in this data set, with voices individuated in this way, there is more allo-repetition than self-repetition for each voice, but that the level is not uniformly greater in the actual dialog than in the randomized dialogs. Only the captain's voice displays a clear difference on both measures in this visualization. The univariate effect sizes from voice and dialog type on mean repetition is shown in Fig. 8: the captain and first officer show less repetition than the other voices recorded (the others individuated as one voice, in this analysis), and the effect of dialog type, with more repetition in actual dialog than in the randomized counterpart, being smaller. The interaction between voice and dialog type for allo-repetition and self-repetition is shown in Fig. 9. The effects of interest in allo-repetition were not significant. In the case of self-repetition, there is significance in the interaction, with the voices of both the captain ($p < 0.05$) and first officer ($p < 0.02$) providing more self-repetition in actual dialog than in the randomizations.[6] The greater difference in repetition between ACTUAL and RANDOMIZED dialogs as measured for self-repetition than for allo-repetition that appears in many other dialog contexts does not exist here.

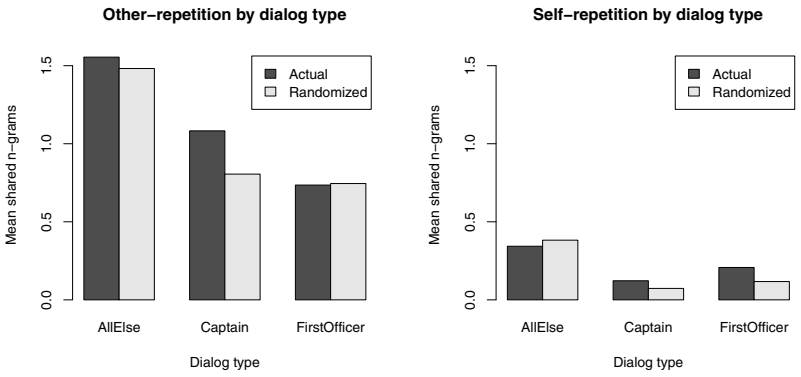


Fig. 7. Flight 1549 speaker means: Random vs Real allo-sharing (L) self-sharing (R)

The temporal overlap effects are shown in Figure 12 (overall (L), analyzing in terms of two distinct speakers vs. all else (M) and by individual channel (R)). Effects depending on DIALOGTYPE=ACTUAL were significant in interaction with the speaker: there was significantly more overlap with other speaking agents in

[6] Using `glm(SelfShared~DialogType*Voice)` and a quasipoisson error family.

the ACTUAL dialogs for both the captain and first officer than for participants generally. In Figure 12 (R), noting that HOT-1 is a channel used by the captain and HOT-2, by the co-pilot, it is clear that propensity to overlap temporally is not even among all participants in this dialog setting. Figure 10 demonstrates the tendency here is for more pronounced allo-repetition than self-repetition for each level of N (although the value of N is not significant as a univariate feature). Figure 11 (left) shows that this varies greatly with each speaker.

Discussion. While repetition is consciously part of the system of air-traffic communication [6], the terseness that is also part of the ritual eliminates other aspects of the language which would ordinarily be open for unconscious repetition (e.g. there is an evident reluctance to use the preposition "to" in discussion of transit towards particular altitudes, lest it be confused with a numeral). By construction, air traffic communication in an emergency situation is not representative of air traffic communication in general. It is an open question how repetition is manifest during air-traffic communications outside crisis events. We have also analyzed this particular corpus with each channel individuated separately, including the separating the voices of the captain and first officer according to intended audience, and point out that strong effects associated with individual channels remain evident. What can be observed is that the pattern of repetition overall, and allo-repetition and self-repetition in isolation, vary with the participant. A generalization supported by this observation is that the participant's role matters as much as the individual filling the role.

4 General Discussion and Final Remarks

For final comparative discussion, we present graphs of the overall sharing of total n-grams for both of the case-studies discussed here in Figures (13)-(14). In both figures, the graph on the left indicates the amount of allo-sharing and the graph on the right shows self-sharing. Within each graph, the bar on the left indicates the levels of sharing in the dialogs as actually ordered, and on the bar on the right indicates the level of sharing in the randomizations. Figure 13 thus shows that in the ATR data, representative of dialogs in which partners show high levels of engagement and mutual interest, there is more self-sharing than allo-sharing, but higher levels of shared tokens in the real data than in the random data on both measures. The Flight 1549 cockpit conversations present rather more allo-repetition is evident than self-repetition. This is consistent with naive expectations of air-flight communications involving much repetition of others to signal understanding.

Of course, it would be useful to explore these dialogs individually in greater depth, rather than dealing with them all in aggregate. We have also begun analysis of the MapTask dialogs using our methods as well [2] and the SwitchBoard dialogs [7,8]. One of our aims is to use the method of quantifying interaction via repetition analysis to be able to assess the extent to which it is possible to reliably quantify the degree of synchrony that exists in a community in relation

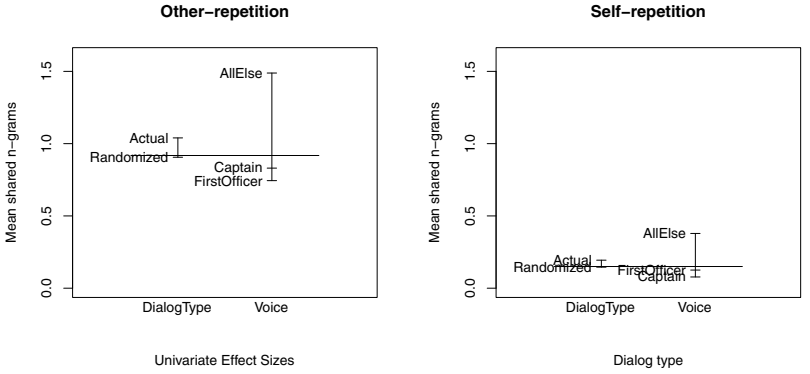


Fig. 8. Flight 1549 effect sizes: Random vs Real allo-sharing (L) self-sharing (R)

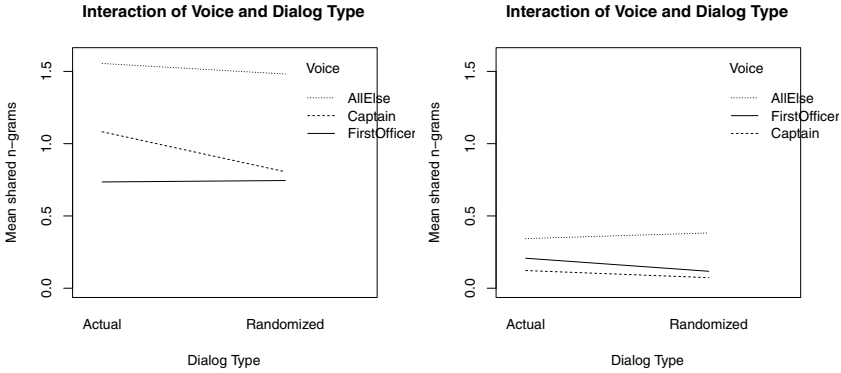


Fig. 9. Flight 1549 interactions: Random vs Real allo-sharing (L) self-sharing (R)

to the origins of synchronization: in some part, it arises through actual interactions; and in other parts, it emerges from agents independently articulating similar points and with the same linguistic expressions. We wish to characterize dialog types and participant role types in relation to patterns of quantified allo-repetition and self-repetition that are evident in the interactions. This has potential application in evaluating systems [16].

One might object to our methods arguing that it is wholly un-natural to consider transcripts instead of underlying audio recordings, or better, full-multimodal corpora such as Campbell [5] collected, or more fruitful to ignore text [3]. Researchers have demonstrated that very rich data sources can be tapped to measure interlocutor involvement in conversation, measuring articulation rates, voice intensity, etc. [10]. While applauding that work, part of our response is that if effects of synchrony can be detected even in the relatively impoverished record of textual transcripts (or interactions that might be recorded in text-based online communities), then it is important exploit this and to refine the means of detection and

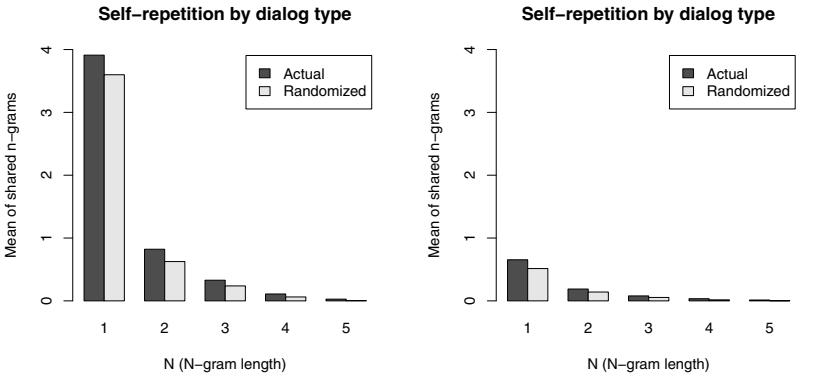


Fig. 10. Dialog type *ACTUAL* vs *RANDOM*: allo-sharing (L) & self-sharing of *N*-grams (R), by N

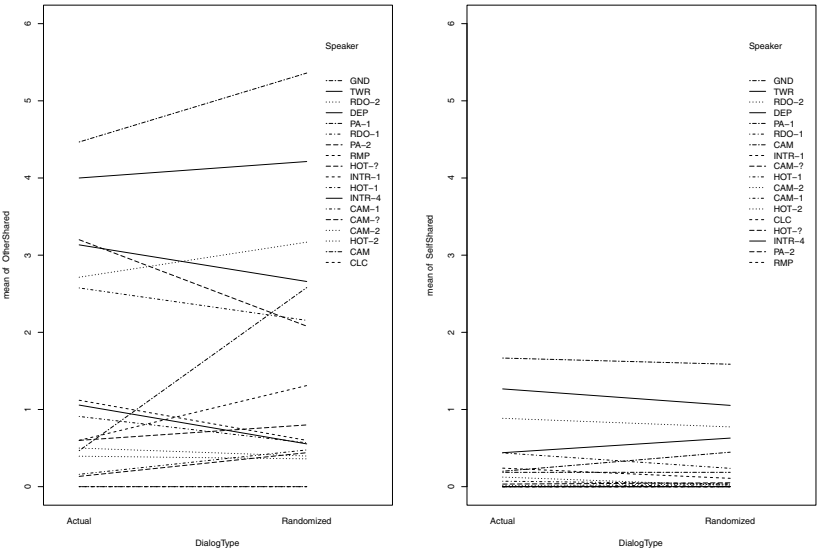


Fig. 11. Interaction of DialogType & Speaker: allo-repetition (left), self-repetition (right)

to establish an interpretative framework for understanding different levels of synchrony. Success implies that the resulting techniques of computational linguistics can contribute to assessment of the naturalness of patently fabricated scripts [9]. Thus, it is necessary further to examine additional dialogs in the light of such an analysis. A different objection is that the measures of temporal overlap used here to contrast with actual overlap are contrived. We would argue that the method is reasonable, but agree that there is more to explore here, including the more

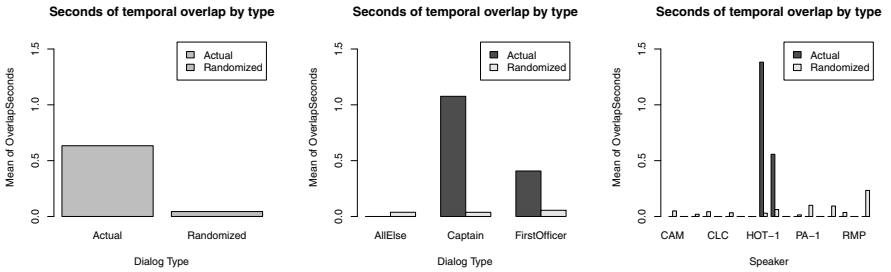


Fig. 12. Temporal overlap: Overall (L), by Speaker (M), by Channel (R)

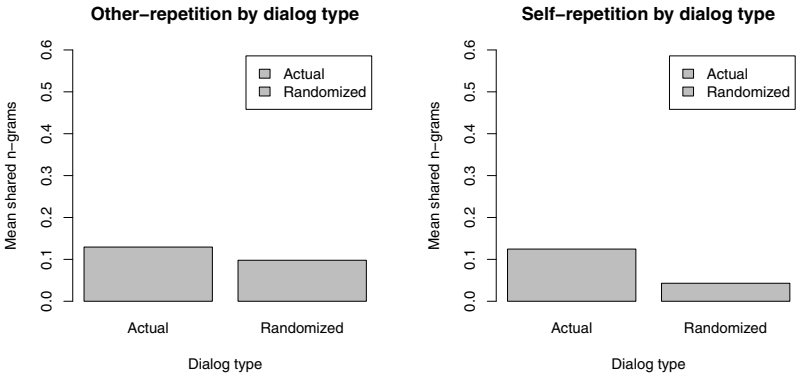


Fig. 13. ATR: Random vs Real allo-sharing (L) self-sharing (R)

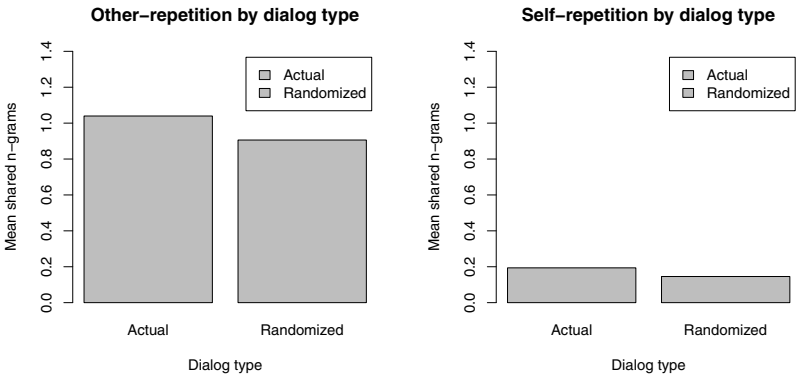


Fig. 14. Flight 1549 group means: Random vs Real allo-sharing (L) self-sharing (R)

sophisticated and quite successful methods recently used by Altmann in analyzing synchronized body motions [1].

The dialogs analyzed may be treated as arbitrary in that they were not recorded with the analysis reported here in mind. Nonetheless, the effects reported in terms of greater overall repetition than with respect to random dialogs, and the effect of greater self-repetition than repetition of others may be integral to the particular data at hand. We think that divergences from this pattern have functional explanations (e.g, a direction giver repeats phrases less than chance would suggest, and personality types of participants matter). Nonetheless, gender, age, educational experience, and all of the other attributes of interlocutors that one might imagine interacting are all left unanalyzed at present.

We currently feel that an overall repetition effect, and more pronounced levels of self-repetition than allo-repetition constitute a signature of synchrony in natural dialog. The self-repetition preponderance (even at $N' = 2+$) may be partly explained by continued maintenance of a dialog plan and partly by the general effects of individual differences in language use that make authorship attribution viable.

Acknowledgements. We are grateful for feedback from anonymous reviewers of an earlier draft of this work for encouragement from Anna Esposito, leader of EU COST Action 2102. Thank you to David Abrahamson for advice on the pragmatics of air-traffic communications.

References

1. Altmann, U.: Investigation of Movement Synchrony Using Windowed Cross-Lagged Regression. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 335–345. Springer, Heidelberg (2011)
2. Bard, E.G., Anderson, A., Sotillo, C., Aylett, M., Doherty-Sneddon, G., Newlands, A.: Controlling the intelligibility of referring expressions in dialogue. Journal of Memory and Language 42, 1–22 (2000)
3. Bouamrane, M.M., Luz, S.: An analytical evaluation of search by content and interaction patterns on multimodal meeting records. Multimedia Systems 13, 89–102 (2007), doi:10.1007/s00530-007-0087-8
4. Brazy, D.: Group chairman's factual report of investigation: Cockpit voice recorder dca09ma026 (2009), docket Number SA-532, Exhibit 12, National Transporation Safety Board, http://ntsb.gov/dockets/aviation/dca09ma026/420526.pdf (last verified June 2010)
5. Campbell, N.: An audio-visual approach to measuring discourse synchrony in multimodal conversation data. In: Proceedings of Interspeech 2009 (2009)
6. Cushing, S.: Fatal Words: Communication Clashes and Aircraft Crashes. University of Chicago Press (1994)
7. Godfrey, J.J., Holliman, E.: Switchboard-1 release 2 (1997), linguistic Data Consortium
8. Graff, D., Bird, S.: Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies. In: Proceedings of the Second International Conference on Language Resources and Evaluation, pp. 427–433. European Language Resources Association, Paris (2000)

9. Murtagh, F., Ganz, A., McKie, S.: The structure of narrative: the case of film scripts. CoRR (2008), http://arxiv.org/abs/0805.3799
10. Oertel, C., De Looze, C., Scherer, S., Windmann, A., Wagner, P., Campbell, N.: Towards the Automatic Detection of Involvement in Conversation. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 163–170. Springer, Heidelberg (2011)
11. Ramseyer, F., Tschacher, W.: Nonverbal Synchrony or Random Coincidence? How to Tell the Difference. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) COST 2102 Int. Training School 2009. LNCS, vol. 5967, pp. 182–196. Springer, Heidelberg (2010)
12. Reitter, D., Keller, F., Moore, J.: Computational modeling of structural priming in dialogue. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pp. 121–124. Association for Computational Linguistics (2006)
13. Reitter, D., Moore, J.: Predicting success in dialogue. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 808–815. Association for Computational Linguistics (2007)
14. Richardson, D.C., Dale, R., Schokley, K.: Synchrony and swing in conversation: coordination, temporal dynamics and communication. In: Wachsmuth, I., Lenzen, M., Knoblich, G. (eds.) Embodied Communication. Oxford University Press (2008)
15. Stormark, K.M., Braarud, H.C.: Infants' sensitivity to social contingency: a "double video" study of face-to-face communication between 2- and 4-month-olds and their mothers. Infant Behavior & Development 27, 195–203 (2004)
16. Sun, X., Nijholt, A.: Multimodal Embodied Mimicry in Interaction. In: Esposito, A., Vinciarelli, K., Vicsi, C., Pelachaud, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 147–153. Springer, Heidelberg (2011)
17. Tannen, D.: Talking voices: repetition, dialogue, and imagery in conversational discourse. Cambridge University Press, Cambridge (2007)

A Companion Technology for Cognitive Technical Systems

Andreas Wendemuth[1] and Susanne Biundo[2]

[1] Cognitive Systems, Otto-von-Guericke University, 39016 Magdeburg, Germany
andreas.wendemuth@ovgu.de
http://www.kognitivesysteme.de
[2] Institute for Artificial Intelligence, University of Ulm, 89069 Ulm, Germany
susanne.biundo@uni-ulm.de
http://www.uni-ulm.de/in/ki/biundo

Abstract. The Transregional Collaborative Research Centre SFB/TRR 62 *"A Companion Technology for Cognitive Technical Systems"*, funded by the German Research Foundation (DFG) at Ulm and Magdeburg sites, deals with the systematic and interdisciplinary study of cognitive abilities and their implementation in technical systems. The properties of multimodality, individuality, adaptability, availability, cooperativeness and trustworthiness are at the focus of the investigation. These characteristics show a new type of interactive device which is not only practical and efficient to operate, but as well agreeable, hence the term "companion". The realisation of such a technology is supported by technical advancement as well as by neurobiological findings. Companion technology has to consider the entire situation of the user, machine, environment and (if applicable) other people or third interacting parties, in current and historical states. This will reflect the mental state of the user, his embeddedness in the task, and how he is situated in the current process.

1 Research Issues

Technical systems of the future are *Companion*-systems - cognitive technical systems, with their functionality completely individually adapted to each user: They are geared to his abilities, preferences, requirements and current needs, and they reflect his situation and emotional state. They are always available, cooperative and trustworthy, and interact with their users as competent and cooperative service partners.

Guided by this vision, in the Transregional Collaborative Research Center SFB/TRR 62, an interdisciplinary consortium of computer scientists, engineers, physicians, neuroscientists and psychologists, are involved with the systematic exploration of cognitive abilities and their realization in technical systems. Here, the properties of individuality, adaptability, availability, cooperativeness and trustworthiness are in the center of the investigation. The aim is to realize these so-called *companion properties* by cognitive processes in technical systems, and

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 89–103, 2012.

to examine them using psychological behavioral models and functional models of brain mechanisms. This will be the foundation for a technology within the realm of affective computing [1] in which human users are offered a completely new dimension in dealing with technical systems. The SFB/TRR 62 complements here the work of other consortia. For example, the EU 6^{th} framework consortium HUMAINE [2] aims at the development of emotion-oriented systems, however a backbone companion architecture was not aimed at. The EU 6^{th} framework COMPANIONS consortium [3] focuses on virtual conversational agents, however neurobiological foundations and planning and strategy aspects were not investigated. Neurobiologically motivated modelling however was investigated by some of the SFB/TRR 62 members and colleagues in the NIMITEK project [4]. These examples are of course non-exhaustive and merely illustrate that other prominent activities with related foci exist.

The Transregional Collaborative Research Centre SFB/TRR 62 was installed at 01. January 2009 by the German Research Foundation (DFG) at sites Ulm and Magdeburg. At the University of Ulm, the Otto-von- Guericke University in Magdeburg, and the Leibniz-Institute for Neurobiology in Magdeburg, 80 scientists are now working within this project.

2 Research Areas

The research program to develop a *Companion* technology for cognitive technical systems includes interdisciplinary research in the methodological basis of the three key areas of *planning and decision-making*, *interaction and availability* and *situation and emotion*. The relevant research questions are investigated from two complementary perspectives on the interaction of user and system:

- The *system perspective* focuses on the structural aspect, i.e. the construction of cognitive functional units and the realization of *companion properties* by these functional units.
- The *user's perspective* examines the effect of system behavior to the user. This manifests itself in the mental model, which the users build up of a *Companion*-system, and in the user's response to the system.

2.1 Planning and Decision Making

The central cognitive processes of planning, reasoning, and decision making are the basis of action and interaction between users and technical systems. To unveil these processes, the development of strategies of action in biological and technical systems are being investigated, as are the effects of system behavior on the user behavior and its effects on the interaction process. The overall goal is to develop, to use and to test knowledge-based methods which enable a *Companion*-system to support users of applications and services in their actions and decisions in a comprehensive, professional and individual manner. A *Companion*-system should be able to provide solutions for complex tasks – all by

itself or in cooperation with the user –, using targeted and traceable strategies, and, in dialog with the user, give decision support and action recommendations. These strategies are based on the user's current interests, on his abilities and preferences; they take into account his emotional state, and adapt the support services to the current environmental situation.

The SFB/TRR 62 deals with the topics of action planning, strategy development and decision finding from both a system and user perspective. To this end, it is quite natural and adequate to rely on a technical equivalent of the user's cognitive abilities, namely on AI planning [5].

Assistance functionalities required for supporting individual users of a technical system include (1) generating a plan of action for a specific user that respects her preferences and emotional state and advising her to carry out the plan in order to achieve a current task, (2) instructing the user on how to escape from a situation where the execution of this plan unexpectedly failed, and (3) justifying and explaining the proposed solution plan in an adequate manner. These assistance functionalities can be provided by an approach to user-centered planning that relies on a domain-independent hybrid planner, a plan repair component, and a plan explanation facility [6].

Plan generation can be performed by reasoning about the preconditions and effects of actions as in partial order causal link (POCL) planning [7]. Hierarchical task network (HTN) planning [8] allows for the use of pre-defined standard solutions and with that enables the exploitation of expert knowledge in solution discovery. By smoothly integrating both paradigms, hybrid planning [9] is particularly well-suited for solving complex real-world planning problems.

When aiming to support human users in a *Companion*-like manner, one of the key aspects is the *individualization* of the recommended course of action. This is achieved by respecting not only the mandatory needs of the person to assist, but also her *user preferences* – solution criteria which are not mandatory, but which should be met if ever possible. Approaches to preference-based planning use (1) temporal constraints on state features as the standard notation for expressing desired, but non-mandatory, plan properties, and (2) heuristic search in the space of states in order to find preferred plans [10]. In contrast to that we have developed an extension to standard POCL techniques that allows to effectively estimate plan quality w.r.t. user preferences [11].

In the context of *Companion*–systems, an additional challenge has to be met, however: Information about the user state, just like information about the world state, depends on sensory input and is thus only partially observable and inherently uncertain. In order to address partial observability, relational partially observable Markov decision processes (POMDPs) can be employed [12]. They allow to handle complex domain dynamics and are thus appropriate to represent decision problems that have to solved when users are to be assisted while operating a technical system. To cope with the accruing uncertainty, we developed a novel approach to hierarchical planning under partial observability in relational domains [13], which combines hierarchical task network planning with the

finite state controller (FSC) policy representation for partially observable Markov decision processes.

Exceptional events can invalidate a plan at execution time. In user-centered planning, a possible remedy is to start from the original plan and try to replace the invalidated parts without touching the unaffected ones, thus ensuring plan stability [14]. But still, having to follow a repaired plan in most cases implies a change of strategy. As this, in general, is a pivotal element of interaction between a *Companion*–system and its user, we investigate the neurophysiological foundations of strategy change in biological systems. We use a suitably designed animal model, which is sufficiently complex to include relevant aspects of strategy change during subject-computer interactions and at the same time sufficiently simple to allow detailed neurophysiological analysis. An important question is, whether and how reinforcement-evaluating brain structures, like the ventral tegmental area (VTA) and lateral habenula (LHb), contribute to a change of behavioral strategy. A change of strategy can be emulated by a contingency reversal of two different acoustic stimuli in a discrimination behavioral task. In order to give an answer to this question we needed to investigate the influence of these brain structures in avoidance learning. A series of experiments were conducted and their results support the general conclusion that VTA and LHb influences the acquisition of an avoidance task and therefore are relevant brain structures for learning strategy changes. Furthermore, it has been confirmed, that stimulation in VTA and LHb has opponent effects on avoidance learning. During acquisition, LHb stimulation impaired avoidance, while VTA stimulation improved it [15]. These neurobiological findings enter into a declarative model that serves to estimate how difficult the management of a particular strategy change would be for a certain individual [16].

Knowledge Base. The knowledge base of a *Companion* system stands at a central point – between perception and action. As such, its task is to incorporate information gained by the sensory system, enhancing it with various kinds of background knowledge, and produce an, as accurate as possible, prediction that can be used as a basis of action selection. The main functional units a knowledge base has to serve are depicted in Figure 1. As can be seen there, information is exchanged with modules working with both symbolic and sub-symbolic data. Perception is rooted in the sub-symbolic processing of audio-visual sensor streams or biosignals taken from respective sensors. The interaction manager maintains communication with the user via multiple modalities and is both a producer of a symbolic stream of observations as well as a consumer of inferred information. The planning system needs to be provided with declarative background and world knowledge.

In order to properly combine declarative expert knowledge with sub-symbolic training data to obtain the required world model, we decided to use Markov Logic [17] as a means to implement knowledge bases of *Companion* systems. So far we have focused on efficiently drawing probabilistic inferences in a dynamic environment while still being able to leverage the expressive power of logic [18]. We also aim at improving sub-symbolic methods for recognizing complex

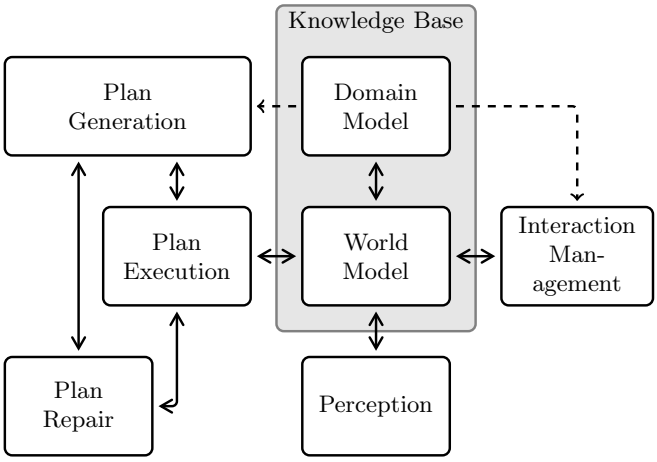


Fig. 1. The Role of the Knowledge Base

dynamic patterns, like activities performed by humans, using a layered Hidden Markov Model architecture [19].

2.2 Interaction and Availability

Humans interact with their environment using all senses, and their cognitive and motor skills. Accordingly a computer-based system, if seen as a peer communication and interaction partner to the human, will also be required to use various input and output channels. Moreover, if the system is going to show *Companion* properties it has to adapt its dialog and interaction strategies and behavior to the environmental situation and to the human's current tasks and emotional state. Dialog and interaction has to be constantly available, perform in a cooperative style and finally should be recognized as a credible, trustworthy interaction of a *Companion*.
Bearing this in mind, it is obvious that future systems showing these *Companion* properties cannot be modeled and developed as current interactive systems are being devised. Interaction needs to be multimodal on multiple devices, situative and individualized, depending on the environmental state and on the human's tasks, habits and emotions. Thus we brought together researchers from neuro biology, from spoken dialog management and from Human–Computer–Interaction computer science in order to understand the requirements in a holistic view and to devise a system approach guided by these expertises.

Neuro-biological Findings. One fundamental rule in interactions is the need for the sender to obtain information that a message has been received, i.e. the subjective sense of completion of an action [20]. In human conversation, language as well as non-verbal means satisfy this expectation [21]. Technical systems are usually not equipped with comparable competences and therefore must

rely more heavily on quick response times to indicate that a user action has been processed. We observed in a functional imaging study that an unexpected delay of feedback by only 500 ms has an equally strong effect on brain activation as a complete omission of the feedback. The increase in activity elicited by delayed and omitted feedback compared with immediate feedback was mainly observed in brain regions known to be involved in attention and action control which suggests that additional neural resources are needed in such potentially irritating situations [22] [23].

Another important aspect in human communication is the fact that it is not only important *what* is said but also *how* it is said, i.e. by means of prosodic modulation. Therefore, we tested the effects of motivational prosody in a learning task by employing short pre-recorded verbal comments (e.g. right, wrong, yes, no) with either neutral or motivational prosody (i.e. praising, blaming). We found that motivational feedback produced a significantly steeper learning curve than feedback with neutral prosody. Additionally, we showed that both of the naturally spoken feedback conditions led to a significantly better learning performance compared with computer-synthesized speech.

Such findings are directly transferred in the project into the technical models, architectures and functionalities as suggested by [24].

Technical Models for Human-Companion Interaction. A *Companion*–system will use context data obtained from diverse sensors via different channels and it will potentially provide interaction with the user through multiple devices using multiple modalities. This multitude of possibilities offers a new perspective for adaptive multimodal interaction and flexible dialog management requiring independent interaction concept for *Companion*–systems as a cross-section technology. Moreover the concept has to provide for run-time adaptivity, because devices, environmental and user statuses define the final user interface only at runtime. To meet this requirement our *Companion* architecture is based on a modified Arch/Slinky meta-model [25]. This Stormy Tree meta-model (Fig. 2) extends the Arch/Slinky model for multimodal interaction. The forked branches on the right side represent different concepts for multimodal interaction. The mediation of different input streams is coordinated by an adaptive fusion process and engine. The multimodal output is coordinated by a rule-based fission engine which selects appropriate information objects according to the goal model being described next.

The hierarchical and modality independent dialog structure is built upon different components (Fig. 3). First of all, goals are the main building blocks of the hierarchy and represent abstract tasks to be accomplished. The top level goals correspond to plan actions from the planning component. For the human-companion interaction process these goals are hierarchically refined using sub-goal links to model parallelism, and are horizontally connected using next-goal links to model sequences. To control goal execution, and enable adaptive dialog strategies, goals can be connected to variables via guards and effects. Guards allow testing for specific values before a goal is executed, while effects represent value assignments to variables when a goal is completed. Each leaf-level goal

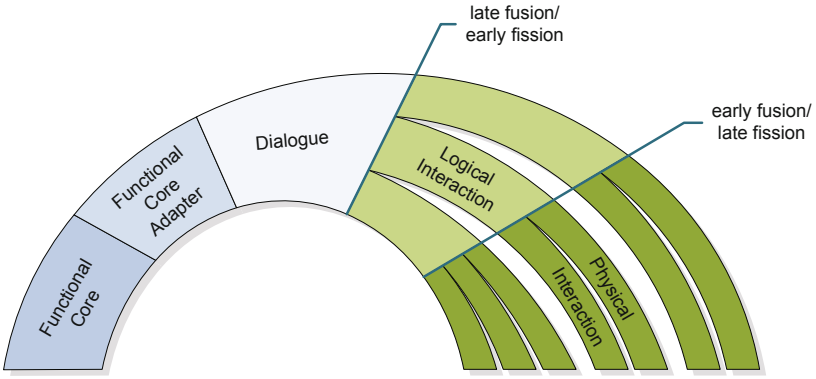


Fig. 2. Stormy Tree Model as an architectural meta model for Human-Companion Interaction

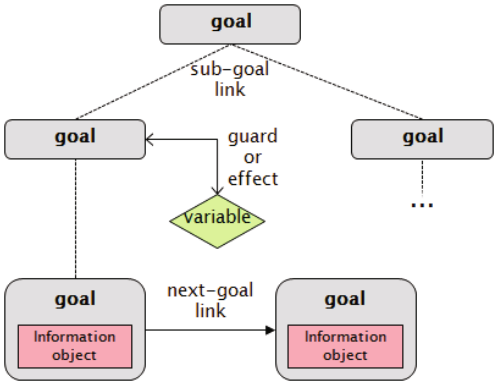


Fig. 3. Goal-based model for dialog and interaction structure

contains information objects as modality independent information, which has to be communicated towards or received from the user.

2.3 Situation and Emotion

Humans can assess situations with emotionally and intentionally acting partners in their entirety and context-dependently. This ability is also to be investigated as key to *Companion*-systems, which makes it another central concern of the SFB/TRR 62. The goal is to obtain a dynamic detection and modeling of the situatedness - location in space, movement, orientation, attention, etc. - and of the individual emotionality in dealing with technical systems. In many typical situations of human-computer interaction it may not be necessary to identify proper emotions (apart from the problem that there seems to be no universally agreed definition of them), instead the system has to identify

the user's disposition towards the current interaction with the system in categories like [26]: engaged/disengaged, frustrated/content, bored/relaxed/under pressure, over-/under-challenged, etc. Using multimodality and fusion [27], [28], high detection rates, robustness and reliability will be reached. The investigation is clustered into a) the dynamic detection and recognition of the environmental situation, intention and emotion of the user, b) modeling aspects and c) interpretation and representation of the overall situation. The need for including information about the emotional state of users into the functionality of a cognitive technical system, results from the neuro-biological fact, that a purely cognitive analysis of human–computer interaction falls short for many areas of human information processing – especially if cognitive technical systems are taking part in social interactions, and when it comes to setting priorities and making decisions. Details of the design of interaction most sensitively account for the success of the sytem-user dialog: In a functional brain imaging study the SFB/TRR 62 observed that an unexpected delay of feedback has a strong effect on brain activation, mainly observed in brain regions known to be involved in attention and action control. This suggests that additional neural resources are needed in such potentially irritating situations. Brain imaging also found that motivational feedback produced a significantly steeper learning curve than feedback with neutral prosody [29]. These findings directly find their way into designing the system's feedback to the user.

Interaction Experiments and Affective Corpora. Because of the importance of emotions in the setting of priorities, in making decisions and controlling actions, the conceptualizing of *Companion*-systems must include situational aspects and emotional processes in dialogs between humans and computers, and it must provide system elements for realization of these effects. Of high importance in this context is the investigation and provision of decision-relevant and actionable corpora within the SFB/TRR 62 from linguistic and non-linguistic human behaviors, which are rarely coded consciously in interpersonal interaction, albeit having great effect on the control of behavior. To that end, we conducted a number of Wizard-of-Oz experiments. The LAST MINUTE experiment allows to investigate interactions of users with a *Companion*–system. It was designed in a way that many aspects of user–companion interaction that are relevant in mundane situations of planning, re-planning and strategy change (e.g. conflicting goals, time pressure, etc.) are experienced by the subjects, with huge number and quality of recorded channels, additional data from psychological questionnaires and semi-structured interviews [30]. In a complimentary Wizard-of- Oz experiment MEMORIZING TASK the emotional load has been induced by a natural language dialog with delay of the commands, non-execution of the command, incorrect speech recognition, offer of technical assistance, lack of technical assistance, and request for termination and positive feedback. This procedure of emotion induction leads the subjects through different locations (octants) in the Valence-Arousal-Dominance (VAD) emotion space.

Four channel peripheral physiological measurements including blood volume pulse (BVP), skin conductance level (SCL) and 2-channel electromyography (EMG) were automatically classified in VAD-space in order to answer the research questions: 1.) to what extent are subject-dependent classifiers of emotion recognition in human computer superior to subject-dependent classifiers and 2.) how robust is the subject-dependent classification transsituational? The study demonstrated that subject-dependent automatical identification of the location in the VAD emotion space outperform significantly with large effect sizes the subject-independent approach in a natural-like human-computer interaction. The data analysis showed that that both EMG signals from corrugators and zygomaticus muscles, HRV and SCL individually differ in their relevance for the classification of emotional responses [31]. However, it remains still unclear whether it will be possible to transsituationally extract stable individual-specific features in different contexts. This may become more possible with sufficient individual–specific information through multimodal assessment with speech prosody, facial behavior and semantic data [27].

Signal Processing for Assessing Situations. In speech, besides sub-symbolic features, also paralinguistic (laughter etc.) and semantic cues play an important role for emotion-expressions. Different emotional analyses tools were applied on different naturalistic corpora, either available or created within our project [32], [33].

In gesture recognition, dynamic (i.e. movement of the hands) and static gestures (postures) are classified [34], taking signs of the American Sign Language (ASL), with recognition rate 94%-99%. For the recognition of dynamic gestures a hidden-markov-model based approach was chosen. Both approaches are currently being integrated [35]. A framework for estimating head poses has been developed [36]. Combining audio information and visual information for the classification of emotional states has been perfomed with multiple features and classifiers [37] on the audio/visual emotion challenge data sets (among top 3 performers, [38]).
Localizing and tracking the involved person(s) is another important issue. Integrating the dependencies among the objects during occlusions is achieved using random finite sets [39].

Some core computational mechanisms of extracting and analyzing nonverbal and visual signals are presented, enabling virtual agents to create socially competent response behaviors. This contributes to installling the basis of social signal processing in companions with human-like abilities [40].

3 Operating Modes of a Companion System

Figure 4 shows the main operational units of a *Companion*-system. A user interacts with a *Companion*-system in a variety of ways and expresses his explicit and implicit requirements of the system usually by multiple, simultaneously

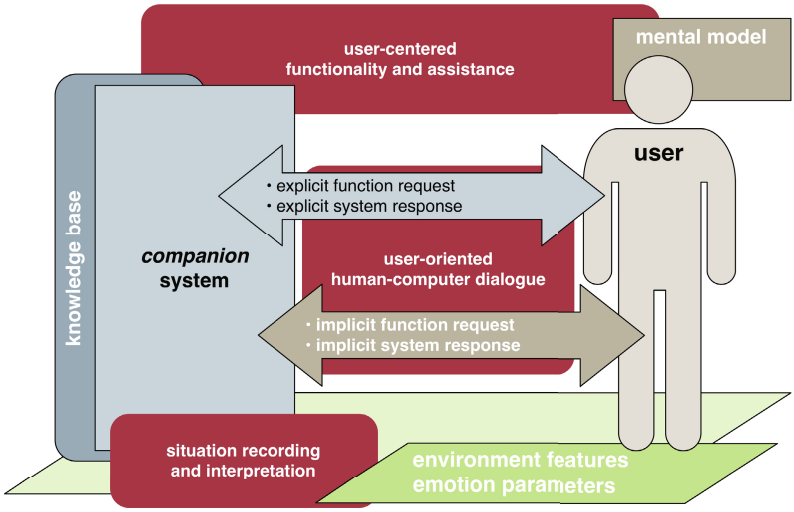


Fig. 4. Operational units of a *Companion*-system

acted, and observable actions. User and system are embedded in an environment that influences both the actions of the user, as well as their observability. A *Companion*-system provides several functions to recognize the intentions of the user, to respond appropriately, and to meet his explicit and implicit requirements. Statements of the user are first recognized as a sequence of observations in different modalities and components, and then classified using operational units for the analysis of language, facial expressions and gestures as well as psychobiological features. The recognition and the assessment of the environment is performed likewise, on the basis of sensory features of the environmental situation. These perceptive features are fused into a sub-symbolic representation of the overall state, representing the emotional state of the user as well as the current environmental situation. The transformation of sub-symbolic representation of the total state into a symbolic state description takes place under inclusion of a stored knowledge model. The symbolic state description provides the basis for intention recognition, and it is utilized for the control of planning and interaction components as well as for dialog control. To realize user-centric functionality and assistance, planning components, based on the current knowledge model, generate action plans and recommendations, and conduct necessary adjustments to the action plans already in progress. The dialog between users and system is performed by components of multimodal interaction. They recognize user actions which are correlated to explicit functional requirements of the *Companion*-system, and which are carried out on different input channels. The associated changes in the state of the application and in the overall situation then result in corresponding system responses. In addition, components of multimodal interaction can realize the output of information to the user, where they use various devices and forms of presentation.

4 Application Perspectives

A *companion technology for cognitive technical systems* is of high relevance for a wide range of applications. Prototypical examples are given below for two application areas.

4.1 Individualized Personal Assistance

Technical solutions for personal assistance systems exist in multitude. A grounded design formalism for such systems, however, is still in research progress [41]. The *Companion*-technology enables the development of a completely new type of personal assistance schemes which are considered reliable companions of their users in carrying out daily tasks and projects in private and professional life. Independent of the location of the user, these assistants are always available, they manage the user's current tasks, they take into account differently prioritized goals, and they support the user in the execution of appropriate actions. The assistants are not only able to react to deviations from planned practices or expected situations in a dynamic and flexible way, but they also independently develop alternative approaches which they pro-actively propose, predicting the consequences. They are familiar with and take into account personal preferences and priorities, they dynamically capture the current environmental situation, as well as the emotional state and the cognitive load of the user, and they accordingly adapt their support services and their communication patterns. They interact with humans, they cooperatively reach solutions, and they may include or opportunistically use specific external services by interacting with third-party systems. The application potential of such assistance systems includes, besides general organizational assistance in professional and private life, especially the support of elder people in their home environment, for which a number of ambient intelligence support scenarios have been identified [42]. In this application, besides assistance in the operation of technical devices or services and personalized support - such as the planning and execution of the daily routine -, even some monitoring functions which can prevent hazards by early detection and if necessary, by requesting assistance from outside, play a central role. The aspects of trust and acceptance of *Companion*-systems are of key importance in this context.

4.2 Medical Assistance Systems

Due to the constant growth of chronic diseases, in the future more and more patients will have to be continuously and individually motivated, supported and guided in the execution of treatment plans, rehabilitation and support measures especially in neurology, geriatrics and pain therapy. This has been clearly stated and identified for various countries by the OECD [43]. Consideration of emotional parameters of patient and therapist is of highest significance for compliance. *Companion*-systems can be utilized here for individualized, interactive

patient information, education and guidance. Interactive information portals designed to cooperate with the patient can then, for example, be configured in such a way that, by individualization regarding the cognitive and perceptual abilities of the individual, a better acceptance of the therapeutic suggestions and actions is achieved. Also conversational medicine (psychotherapy, psychosomatic medicine and psychiatry, rehabilitation psychology), which is already heavily based on linguistic communication, increasingly uses information technologies to support their therapies [44], which can benefit greatly from the concepts of individuality, adaptability and trustworthiness. Increasingly important, and well advanced in many countries, is tele-medical care. The treatment of chronically ill patients, old patients, rehabilitation patients and patients in special situations (communication with aircraft, ship and spacecraft crews during the illness of a passenger, assistance of medical personnel in crisis missions, etc.) can be supported by *Companion*-systems to a considerable extent. In connection with medical applications, which increasingly assume an active cooperation of the patient, trustworthiness and acceptance of the systems is of substantial importance. Optimal compliance also requires that the patient experiences the system as individually tailored to his needs, adaptable, universally available and empathic, and that on this basis the patient confidentially interacts with the system.

5 Outlook

Detailed information on subprojects of the SFB/TRR 62 as well as recent publications and research results are available at http://www.sfb-trr-62.de/.

Acknowledgments. The *Companion Technology* is being developed within the Transregional Collaborative Research Centre SFB/TRR 62 *"A Companion Technology for Cognitive Technical Systems"*, funded by the German Research Foundation (DFG). All concepts and joint research activities presented in this paper result from the collaboration of the principal investigators A. Al-Hamadi, S. Biundo, A. Brechmann, K. Dietmayer, J. Frommer, H. Kessler, B. Michaelis, W. Minker, H. Neumann, F.W. Ohl, G. Palm, D. Rösner, H. Scheich, F. Schwenker, H. Traue, M. Weber, A. Wendemuth.

References

1. Picard, R.: Affective Computing. The MIT Press, Cambridge (2007) ISBN 0-262-16170-2
2. Cowie, R. (coordinator): EU-IST Network of Excellence HUMAINE (The Human Machine Interaction Network on Emotion) (2004-2007), emotion-research.net
3. Wilks, Y. (coordinator): EU-IST Integrated Project IST-34434 COMPANIONS (2006-2010), companions-project.org
4. Wendemuth, A., Braun, J., Michaelis, B., Ohl, F., Rösner, D., Scheich, H., Warnemünde, R.: Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems (NIMITEK). In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) PIT 2008. LNCS (LNAI), vol. 5078, pp. 141–144. Springer, Heidelberg (2008)

5. Ghallab, M., Nau, D., Traverso, P.: Automated Planning: Theory & Practice. Morgan Kaufmann Publishers Inc., San Francisco (2004) ISBN 1558608567
6. Biundo, S., Bercher, P., Geier, T., Müller, F., Schattenberg, B.: Advanced user assistance based on AI planning. Cognitive Systems Research 12(3-4), 219–236 (2011)
7. Penberthy, J., Weld, D.: UCPOP: A Sound, Complete, Partial Order Planner for ADL. In: Proceedings of the Third International Conference on Knowledge Representation and Reasoning, pp. 103–114 (1992)
8. Nau, D., Au, T., Ilghami, O., Kuter, U., Muñoz-Avila, H., Murdock, J., Wu, D., Yaman, F.: Applications of SHOP and SHOP2. IEEE Intelligent Systems (2004)
9. Biundo, S., Schattenberg, B.: From Abstract Crisis to Concrete Relief (A Preliminary Report on Combining State Abstraction and HTN Planning). In: Proceedings of the 6th European Conference on Planning (ECP 2001), pp. 157–168. Springer (2001)
10. Coles, A., Coles, A.: LPRPG-P: Relaxed Plan Heuristics for Planning with Preferences. In: Proceedings of the 21st International Conference on Automated Planning and Scheduling (ICAPS 2011), pp. 26–33 (2011)
11. Bercher, P., Biundo, S.: Hybrid Planning with Preferences Using a Heuristic for Partially Ordered Plans. In: 26th PuK Workshop "Planen, Scheduling und Konfigurieren, Entwerfen", PuK 2011 (2011)
12. Sanner, S., Kersting, K.: Symbolic dynamic programming for first-order POMDPs. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010), pp. 1140–1146 (2010)
13. Müller, F., Biundo, S.: HTN-Style Planning in Relational POMDPs Using First-Order FSCs. In: Bach, J., Edelkamp, S. (eds.) KI 2011. LNCS, vol. 7006, pp. 216–227. Springer, Heidelberg (2011)
14. Bidot, J., Schattenberg, B., Biundo, S.: Plan Repair in Hybrid Planning. In: Dengel, A.R., Berns, K., Breuel, T.M., Bomarius, F., Roth-Berghofer, T.R. (eds.) KI 2008. LNCS (LNAI), vol. 5243, pp. 169–176. Springer, Heidelberg (2008)
15. Ilango, A., Shumake, J., Wetzel, W., Scheich, H., Ohl, F.: Effects of ventral tegmental area stimulation on the acquisition and long-term retention of active avoidance learning. Behav. Brain Res. 225(2), 515–521 (2011)
16. Schulz, A., Schattenberg, B., Woldeit, M., Brechmann, A., Biundo, S., Ohl, F.W.: Reinforcement learning and planning models for two-way-avoidance and reversal learning. In: Proc. Annual Meeting of the Society For Neuroscience, Washington, USA (2011)
17. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62(1-2), 107–136 (2006)
18. Geier, T., Biundo, S.: Approximate Online Inference for Dynamic Markov Logic Networks. In: Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence, pp. 764–768 (2011)
19. Glodek, M., Bigalke, L., Palm, G., Schwenker, F.: Recognizing Human Activities Using a Layered HMM Architecture. Machine Learning Reports 5, 38–41 (2011)
20. Miller, R.B.: Response time in man-computer conversational transactions. In: Proceedings AFIPS Spring Joint Computer Conference, Montvale, pp. 267–277 (1968)
21. Clark, H.H., Brenan, S.E.: Grounding in communication. In: Resnick, L.B., Levine, J.M., Behrend, S.D. (eds.) Perspectives on Socially Shared Cognition, 1st edn., pp. 127–149. Amer Psychological Assn., Washington (1991)
22. Kohrs, C., Behne, N., Scheich, H., Brechmann, A.: Similiar fMRI activation by delayed and omitted visual feedback. In: Proceedings of the Annual Meeting of the Society For Neuroscience, Chicago, USA (2009)

23. Kohrs, C., Angenstein, N., Scheich, H., Brechmann, A.: The temporal contingency of feedback: effects on brain activity. In: Proceedings of the International Conference on Aging and Cognition, Dortmund, Germany (2010)
24. Oviatt, S.: Multimodal Interfaces. In: Sears, A., Jacko, J. (eds.) The Human-Computer Interaction Handbook, 2nd edn., pp. 413–432. CRC Press, Boca Raton (2008)
25. Gram, C., Cockton, G.: Design principles for interactive software. Chapman & Hall, Ltd., London (1997) ISBN 0-412-72470-7
26. Scherer, S., Schels, M., Palm, G.: How Low Level Observations Can Help to Reveal the User's State in HCI. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 81–90. Springer, Heidelberg (2011)
27. Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C., Schwenker, F.: Multimodal Emotion Classification in Naturalistic User Behavior. In: Jacko, J.A. (ed.) HCII 2011, Part III. LNCS, vol. 6763, pp. 603–611. Springer, Heidelberg (2011)
28. Glodek, M., Scherer, S., Schwenker, F., Palm, G.: Conditioned Hidden Markov Model Fusion for Multimodal Classification. In: ISCA (publ.): Proceedings of Interspeech 2011, pp. 2269–2272 (2011)
29. Wolff, S., Kohrs, C., Scheich, H., Brechmann, A.: Temporal contingency and prosodic modulation of feedback in human-computer interaction: Effects on brain activation and performance in cognitive tasks. In: Heiss, H., Pepper, P., Schlingloff, H., Schneider, J.(eds.): Informatik 2011 - Informatik schafft Communities: 41. Jahrestagung der GI, 4.-7.10. LNI P-192. Springer, Berlin (2011)
30. Rösner, D., Friesen, R., Otto, M., Lange, J., Haase, M., Frommer, J.: Intentionality in Interacting with Companion Systems – An Empirical Approach. In: Jacko, J.A. (ed.) HCII 2011, Part III. LNCS, vol. 6763, pp. 593–602. Springer, Heidelberg (2011)
31. Tan, J., Walter, S., Scheck, A., Hrabal, D., Hoffmann, H., Kessler, H., Traue, H.: Repeatability of facial electromygraphy (EMG) activity over corrugator supercilii and zygomaticus major on differentiating various emotions. Journal of Ambient Intelligence and Humanized Computing (2011) online
32. Siegert, I., Böck, R., Philippou-Hübner, D., Vlasenko, B., Wendemuth, A.: Approriate emotional Labelling of non-acted speech using basic emotions, Geneva emotion wheel and self-assessment maninkins. In: Proceedings of the 2011 IEEE International Conference on Multimedia & Expo. (ICME 2011), Barcelona, Spain, July 11-15, pp. 1–6 (2011)
33. Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting Laughter in naturalistic multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. To appear in ACM Transactions on Interactive Intelligent Systems: Special Issue on Affective Interaction in Natural Environments (2011)
34. Al-Hamadi, A., Rashid, O., Michaelis, B.: Posture Recognition using Combined Statistical an Geometrical Feature Vectors based on SVM. International Journal of Information and Mathematical Sciences 6, 7–14 (2010)
35. Rashid, O., Al-Hamadi, A., Michaelis, B.: Integration of Gesture and Posture Recognition Systems for Interpreting Dynamic Meanings using Particle Filter. In: International Conference on Soft Computing and Pattern Recognition, Paris, pp. 47–50 (2010)

36. Layher, G., Liebau, H., Niese, R., Al-Hamadi, A., Michaelis, B., Neumann, H.: Robust Stereoscopic Head Pose Estimation in Human-Computer Interaction and a Unified Evaluation Framework. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part I. LNCS, vol. 6978, pp. 227–236. Springer, Heidelberg (2011)
37. Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., Schwenker, F.: Multiple Classifier Systems for the Classification of Audio-Visual Emotional States. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 359–368. Springer, Heidelberg (2011)
38. Schuller, B., Valstar, M., Cowie, R., Pantic, M.: The First Audio/Visual Emotion Challenge and Workshop – An Introduction. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, p. 322. Springer, Heidelberg (2011)
39. Reuter, S., Dietmayer, K.: Pedestrian Tracking Using Random Finite Sets. In: 14. International Conference on Information Fusion, Chicago, pp. 1–8 (2011) ISBN: 978-1-4577-0267-9
40. Layher, G., Tschechne, S., Scherer, S., Brosch, T., Curio, C., Neumann, H.: Social Signal Processing in Companion Systems - Challenges Ahead. In: Heiss, H., Pepper, P., Schlingloff, H., Schneider, J.(eds.): Informatik 2011 - Informatik schafft Communities: 41. Jahrestagung der GI, 4.-7.10. LNI P-192. Springer, Berlin (2011)
41. Goldkuhl, G., Lind, M.: A Multi-Grounded Design Research Process. In: Winter, R., Zhao, J.L., Aier, S. (eds.) DESRIST 2010. LNCS, vol. 6105, pp. 45–60. Springer, Heidelberg (2010)
42. van Hoof, J., Kort, H.S.M., Rutten, P., Duijnstee, M.: Ageing-in-place with the use of ambient intelligence technology: Perspectives of older users. International Journal of Medical Informatics 80(5), 310–331 (2011)
43. Organization for Economic Cooperation and Development (OECD) (publ.): Long-Term Care for Older People: The OECD Health Project. OECD Publishing, Paris (2005) ISBN: 92-64-00848-9
44. Scherer, M.: Rehabilitation psychology. Corsini Encyclopedia of Psychology, pp. 1-3. Wiley Online Library (2010)

Semantic Dialogue Modeling

Günther Wirsching[1], Markus Huber[2], Christian Kölbl[2],
Robert Lorenz[2], and Ronald Römer[3]

[1] Katholische Universität Eichstätt-Ingolstadt, Math.-Geogr. Fakultät
guenther.wirsching@ku-eichstaett.de
[2] Universität Augsburg, Institut für Informatik
{markus.huber,christian.koelbl,robert.lorenz}@informatik.uni-augsburg.de
[3] Brandenburgische Technische Universität Cottbus, Fakultät 3
ronald.roemer@tu-cottbus.de

Abstract. This paper describes an abstract model for the semantic level of a dialogue system. We introduce mathematical structures which make it possible to design a semantic-driven dialogue system. We describe essential parts of such a system, which comprise the construction of feature-values relations representing meaning from a given world model, the modeling of the flow of information between the dialogue strategy controller and speech recogniser by a *horizon of comprehension* and the *horizon of recognition results*, the connection of these horizons to wordings via *utterance-meaning pairs*, and the incorporation of new horizons into a state of information. Finally, the connection to dialogue strategy controlling is sketched.

Keywords: entity-relationship, weighted feature-values relation, semantic representation, utterance-meaning pairs, dialogue modeling.

1 Introduction

This paper describes an abstract model for the semantic level of a dialogue system. The task of such a system is to collect the data needed to perform certain actions. Technically, this can be described as extraction, insertion, deletion, and change, of entries in a database. We model the information available to the system using the mathematical notion *weighted feature-values relation*. The feature-values relation flowing through the system are algorithmically derived from a world model containing data and actions, where the data is given via an SQL database and an appropriate entity-relationship (abbreviated ER) diagram.

The connection between the semantic level and an automatic speech recogniser is given by *utterance-meaning pairs*, which we motivate by a model stemming from behavioristic psychology. Utterance-meaning pairs associate feature-values relations representing meaning to possible wordings expressing a meaning, and vice versa. Technically, the association from wordings to meanings can be realised by a weighted finite state transducer, which we call the *UMP-transducer*. This chaining of a representation of semantic by a feature-values relation on one hand,

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 104–113, 2012.

and a language model describing possible wordings to express the meanings on the other hand, allows the design of a semantic-driven dialogue system.

On the semantic level, we store the background information of the system in a *state of information*, which is also a weighted feature values relation. The flow of information between the semantic level and a speech recogniser is given by dynamically generated *horizons* which contain, in each situation, the meanings which may play a role in the given situation. In a given dialogue turn, when a user input is expected, a *horizon of comprehension* is sent to the recogniser. Using utterance-meaning pairs, the recogniser is able to construct dynamically an appropriate language model which can be used for recognition. The recognition results are sent to the UMP-transducer, which converts them into a *horizon of recognition results*, also represented as weigthed feature-values relation. Now the task of the dialogue strategy controller is to incorporate the horizon of recognition results into the state of information, and to decide what to do next, based on the now available information.

The paper starts with a description of the world model, a formal definition of feature-values relation, and an indication how our algorithm constructs feature-values relations from the world model. The flow of information is illustrated by an example dialogue, followed by an introduction to utterance-meaning pairs, and a description how to construct a horizon of comprehension in a given situation. Finally, it is indicated how the dialogue strategy controller has to deal with the state of information and the horizons.

2 World Model and Feature-Values Relation

Our point of departure is a world model consisting of two parts: a set of data, and a set of possible actions. For definiteness and simplicity, we assume that the data is given via an SQL-database, together with an appropriate ER-diagram, but we emphasize that the structures which we use in the sequel can also be derived from other data structures. With respect to the action, we assume that a list of possible action is given, where to each action, possible sets of data needed to perform the action are specified.

We use *feature-values relations* as the mathematical structure carrying semantic information. Here is a formal definition:

Definition 1. *A* feature-values relation *(FVR) is a finite acyclic labeled directed graph $R = (V, \rightarrow, \ell)$, where*

- *V is a finite set of labels,*
- *$\rightarrow \subset V \times V$ is an acyclic relation,*
- *$\ell : V \rightarrow L$ is a labeling of vertices, where L is a set of labels.*

If an FVR is given, an *initial vertex* is, by definition, a vertex without incoming arrow, and a *terminal vertex* is one without outgoing arrow.

In [2], an algorithm which transforms a pair consisting of an SQL-database and an appropriate ER-diagram into a feature-values relation is described. The

algorithm starts by constructing, to each given entity type, an *elementary* FVR modeling the attributes and relations of the given entity type. Moreover it contains structures which we call *anchors*, where each anchor corresponds to an entity type which is involved in some relation with the given entity type, and to its role in the relation.

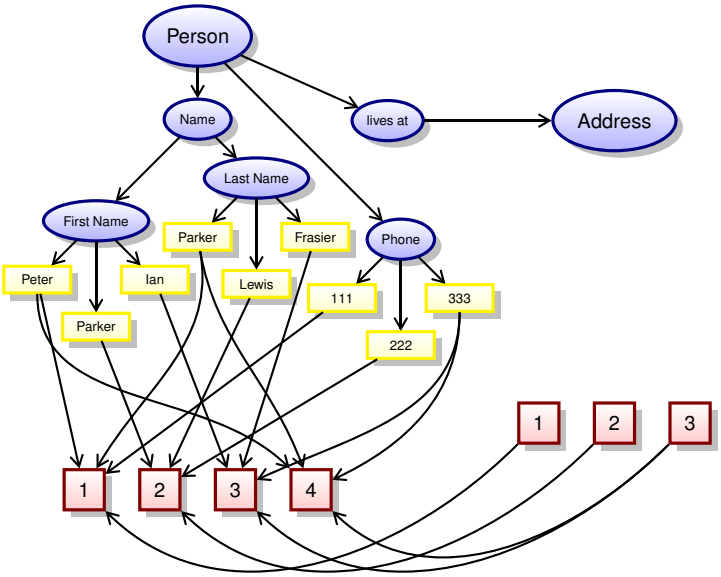


Fig. 1. The elementary feature-values relation associated to entity type "Person", with an anchor associated to entity type "Address". The ID-layer of the anchor is connected the ID-layer of the initial FVR according to the relation given by the database.

In our simple example, the given entity type is "Person" with attributes "First Name", "Last Name", and "Phone", and a relation "lives at" connecting each person to an entity of type "Address". The constructed elementary FVR uses the given entity type "Person" as *root feature*, which is an initial vertex in the elementary FVR, and an *ID-layer* consisting of a set of terminal vertices. By construction, there is a one-one-correspondance between IDs in the ID-layer and entities of the given type in our SQL-database.

In figure 1, there is also an anchor: it consists of a terminal vertex labeled "Address", which is reachable from the vertex labeled "Person" via the relation "lives at". an a set of initial vertices corresponding to the IDs of entities of type "Address" in the SQL-database. The anchor can be thought of as a placeholder for the elementary FVR constructed from the entity type "Address".

Starting with an elementary FVR and putting copies of elementary FVRs, as far as needed, in appropriate anchors, we have a recursive construction of FVRs

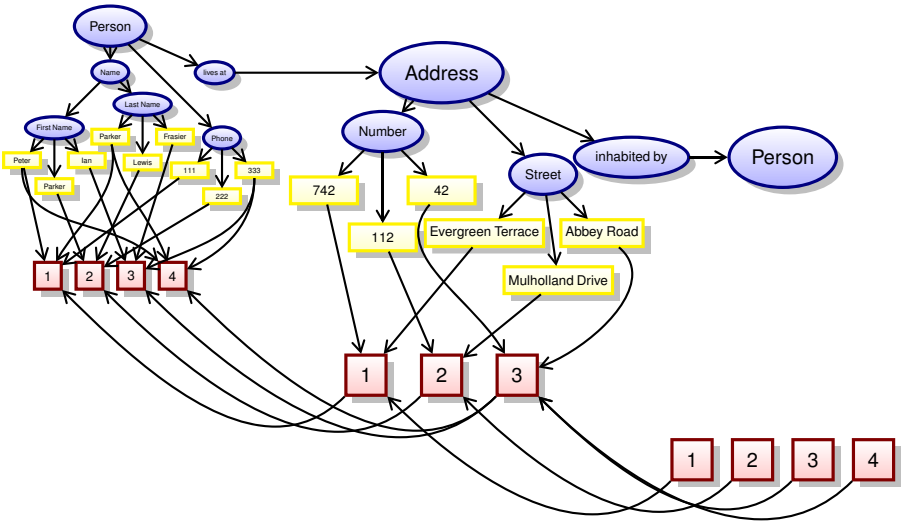


Fig. 2. Recursive construction of an FVR to entity type "Person", with the anchor associated to entity type "Address" filled by the corresponding elementary FVR

Fig. 3. A set of possible actions, represented as FVR

from an ER-diagram and an appropriate SQL-database. The recursion depth is, in principle, arbitrary (but finite).

The set of actions can also be represented as FVR, as is indicated in figure 3.

3 An Example Dialogue

USER: I want to call Parker.
SYSTEM: Is Parker the first name?
USER: No, I mean Peter Parker.
SYSTEM: Which Peter Parker do you want to call?
USER: Change that terrible song to something from Johnny Cash.
SYSTEM: Which album by Johnny Cash?
USER: At San Quentin
⟨`system starts playing`⟩

SYSTEM: Which Peter Parker do you want to call?
USER: The one who lives at 742 Evergreen Terrace.
⟨`calling the selected partner`⟩

4 Meaning and Utterance

The control of a dialogue system relies on the *meanings* of what the user says. The system has to gather those pieces of information which are needed for performing a specific task but are not yet given by the user. Which means it has to ask for it. The following properties will help to clarify our ideas how we model *meanings* flowing through a dialogue system:

- Important for dialogue control are the *meanings* of each utterance, not the precise wording.
- *Meaning* can be represented by a feature-values relation.
- *Meaning* is conveyed by an *utterance*.

In order to get an idea how *meaning* is connected to an *utterance*, we have a look on an idea from psychology. Skinner [4] applies the formal scheme, crucial for behaviorism,

$$\text{Stimulus} \longrightarrow \text{Response} \longrightarrow \text{Consequences}$$

to "verbal behavior" as follows:

Stimulus: the context of a verbal behavior,
Response: the utterance itself,
Consequences: possible impacts in the given context.

Moreover, he asserts that *meaning*

- is not a property of the utterance,
- is to be constructed from context and consequences.

In these terms, the ideal aim of behavioristic psychology is to describe, given stimulus and consequences, the set of possible fitting utterances, together with a probability distribution on this set. If this aim could be reached, it would also be perfect for the speech recognition task in dialogue modeling: a given probability distribution on a given set of utterances can be transformed into a language model apt for configuring a speech recognizer. The language model would be optimal for the speech recognition task, if it represents the 'true' probability distribution of utterances in the given situation defined by stimulus and consequences.

With this language modeling aim in mind, a formalization of the Stimulus-Response-Consequences scheme into a mathematical concept "utterance-meaning pair" is described in [8]. Here we note just the definition:

Definition 2. *An* utterance-meaning pair *consists of an* utterance*, described as a word sequence, and a* meaning*, given by a feature-values relation.*

Note that, at this stage, we do not specify the way in which the word sequence describing the utterance is given to the system. In fact, there are different possibilities:

1. As a sequence of words in usual graphemic notation.
2. A phonetic transliteration, taking into account possible slurring of words, or other phonetic variations.
3. Either of the above, enriched by additional prosodic and/or dynamic information.

Moreover, note that the relation "utterance $\leftrightarrow$ meaning" usually is many-to-many:

- Two different utterances may have the same meaning.
- One utterance may have more than one meaning.

Example 1. Here is a simple example of an utterance with two possible meanings:

- Utterance: "I want to *call Parker*"
- Meaning 1: Action = Call, First Name = Parker.
- Meaning 2: Action = Call, Last Name = Parker.

Utterance-meaning pairs are the "atoms" for *semantic dialogue modelling*. In functional regard, which is the important one for dialogue modelling, we may always view the set of utterance-meaning pairs as a mathematical relation, i. e., as a subset of the cartesian product of a set of possible utterances with a set of possible meanings. But it is generally not necessary to store the needed utterance-meaning pairs in a large list. In many cases, it is preferable to define them implicitly in a grammar, and to use a finite state transducer (abbreviated FST) to configure a speech recognizer with a set of utterance-meaning pairs. In addition, the FST may be enhanced with weights, the computation of which is, ideally, based on statistical data from language observations.

5 Horizon of Comprehension

In semantic dialogue modelling, we view a speech recogniser as a black box with three input channels and one output channel, as depicted in figure 4.

In this setting, the *horizon of comprehension* is to be given as a set of possible meanings. As before, there is no need to store this as a (possibly large) set of meanings; it suffices to have an implicit algorithmic description enabling the system to construct this set. The horizon of comprehension can also be endowed with a weight for each meaning. These weights should, for instance, represent *a priori* knowledge about which meaning the user is more (or less) likely to use. If possible, the weights can be chosen to encode Bayesian prior probabilities to each possible meaning.

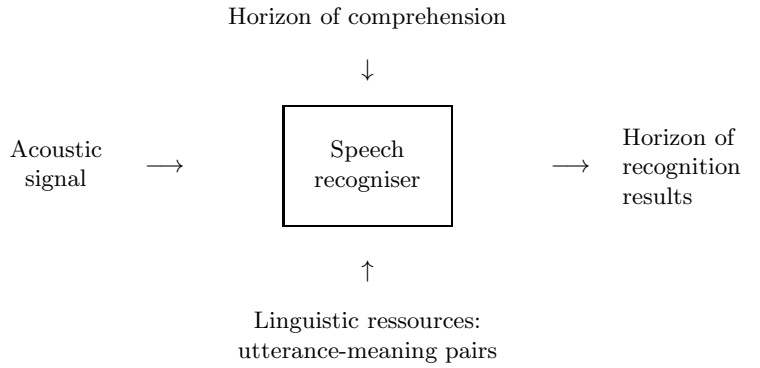


Fig. 4. Configuration of the speech recogniser in semantic dialogue modeling

At each dialogue turn, the speech recogniser is to be given all meanings which should be understandable in the actual context. In a given context (a given dialogue turn), the horizon of comprehension is just the set of meanings which should be understandable in this context. As described in [8], this set can be divided into five parts:

$\mathcal{E}$ **(Horizon of Expectation):**
Set of meanings exactly asked for by the prompt.
$\mathcal{U}$ **(Underanswering):**
Set of meanings answering the prompt only partially.
$\mathcal{O}$ **(Overanswering):**
Set of meanings containing more information than asked for by the prompt.
$\mathcal{D}$ **(Deviating answer):**
Set of meanings overanswering part of what has been asked for.
$\mathcal{G}$ **(Generally available meanings):**
Set of generally available meanings,
e.g., aborting or interrupting the current task.

Each of these sets is a set of meanings. Having, in the background, a set UMP of given utterance-meaning pairs, UMP defines a map associating to each given set $\mathcal{M}$ of meanings a set $w(\mathcal{M})$ of utterances u with the property that there is a meaning $m \in \mathcal{M}$ such that $(u, m) \in \text{UMP}$.

Example 2. Let us consider a context defined by

USER: "I want to call Parker."
SYSTEM: "Is Parker the first name?"

Now the system is waiting for an answer, and the speech recogniser should be configured in a way enabling it to understand any reasonable answer. Here are some examples of utterances for the different parts of the horizon of comprehension:

"No, Parker is the last name." $\in w(\mathcal{E})$
"I don't know." $\in w(\mathcal{U})$
"No, I mean Peter Parker." $\in w(\mathcal{O})$
"I don't know, but he lives at 742 Evergreen Terrace." $\in w(\mathcal{D})$
"Abort calling Parker." $\in w(\mathcal{G})$
"Change that terrible song to something from Johnny Cash." $\in w(\mathcal{G})$

Now we are ready to explain figure 4 more specifically.

- The *horizon of comprehension* is, clearly, context-dependent; it changes from dialogue turn to dialogue turn. In each situation, it depends on the most recent system prompt, on information which the system had received previously, and on the general context of the dialogue which includes all possible executable actions.
- The *linguistic resources* have to be structured in such a way that, for a given set $\mathcal{M}$ of meanings, the set of "wordings" $w(\mathcal{M})$ is easily accessible. Ideally, these sets are endowed with weights for each wording, which can be combined with weights from the horizon of comprehension to give a probabilistic language model for the recogniser.
- The *horizon of recognition results* (abbreviated HoRs) is a weighted set of possible meanings, where the weights are computed from recognition scores. A method for this computation is given in [6]. Note that we don't need the precise wordings of the recognition results, dialogue control works exclusively with meanings. In figure 4, we understand that parsing is included in recognition.

6 The State of Information

On the semantic level, the necessary information is stored in a *State of Information* (abbreviated SoIn). The mathematical structure of the SoIn is weighted FVR, where the weights, and, if necessary, also the structure are changed during the dialogue. An example for a SoIn is given in figure 5.

6.1 Storing Information

The first task of the global SoIn is to store the information collected by extracting possible meanings from the utterances of the user.

Parallel Worlds. On the uppermost level, the global SoIn is a set of unconnected conflicting SoIns representing *parallel worlds*. Each parallel world is equipped with a confidence indicating how sure the system is that this world is what the user intended.

Concentrating on a specific parallel world, the next level consists of a stack of sub-Soins, where each sub-SoIn corresponds to a possible topic.

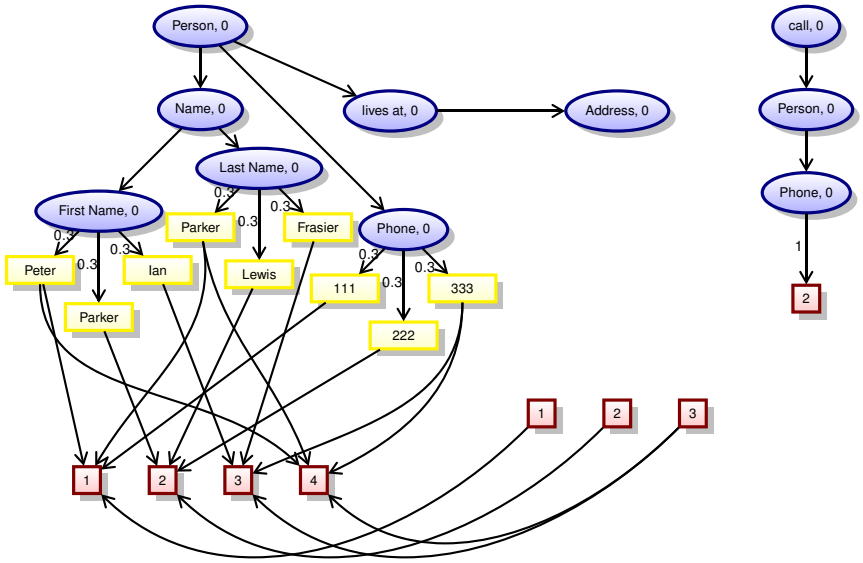


Fig. 5. A State of Information

Mathematical Structure Essentially, each sub-SoIn consists of two parts:

1. The *action part* incorporates identifiers for the possible actions the system is able to perform. Each possible action is equipped with a weight giving an estimate for the probability that the user intends this action. At each dialogue turn, these estimates are to be updated from the weights of the meanings understood by the speech recogniser.
2. The *data part* containing the references to data from the database. It is a weighted FVR, where the weights are appropriately initialized and updated at each dialogue turn. The FVR representing the data part of a sub-SoIn is built recursively from elementary FVRs extracted from the ER-diagram and the data. At each dialogue turn, both the recursion depth and the weights depend on the dialogue history up to that turn.

6.2 The Update Process

The update of the SoIn after an utterance from the user was processed by the speech recogniser is based on the result of this processing, the HoRs, which again is a weighted FVR.

Initially, the HoRs is a list of meanings, where each meaning comes with a score representing its *Bayesian a posterior* probability. (Here the recognition is modeled as Bayesian update process where the prior is given by the language model and the result is the posterior.) The HoRs is separated into sets of meanings with common feature structure, where each feature structure corresponds to

a parallel world; see [6] for more details. Then each set of meanings belonging to one parallel world is incorporated into the appropriate parallel world. According to [3] all involved weighted FVRs can be converted to weighted FSTs and the update can be computed by FST-algorithms.

6.3 The Dialogue Strategy Controller

For the time being, we consider the dialogue strategy controller (abbreviated DiSCo) as a black box with the following specification:

Input: the old SoIn plus the updated SoIn.
Output: a new SoIn plus a horizon of comprehension for the next dialogue turn.
Task: apply strategies to disambiguate collected meanings, and decide what is the next piece of information to be asked for.

References

1. Huber, M., Kölbl, C., Lorenz, R., Wirsching, G.: Ein Petrinetz-Modell zur Informationsübertragung per Dialog. In: Proceedings of the 15th German Workshop on Algorithms and Tools for Petri Nets, AWPN 2008, Rostock, Germany, September 26-27, pp. 15–24 (2008)
2. Huber, M., Kölbl, C., Lorenz, R., Römer, R., Wirsching, G.: Semantische Dialogverarbeitung mit gewichteten Merkmal-Werte-Relationen. In: Hoffmann, R. (Hrsg.) Elektronische Sprachsignalverarbeitung 2009, Tagungsband der 20. Konferenz, Dresden, 21. bis 24. Studientexte zur Sprachkommunikation, vol. 54, pp. S.25–S.32 (September 2009)
3. Kölbl, C., Huber, M., Wirsching, G.: Endliche gewichtete Transduktoren als semantischer Träger. In: Kröger, B.J., Birkholz, P. (Hrsg..) Elektronische Sprachsignalverarbeitung 2011, Tagungsband der 22. Konferenz, Aachen, 28. bis 30. Studientexte zur Sprachkommunikation, vol. 61, pp. S.176–S.183 (September 2011)
4. Skinner, B.F.: Verbal Behavior. Prentice Hall, Englewood Cliffs (1957)
5. Wirsching, G., Huber, M.: Kölbl, C.: The confidence-probability semiring. Technischer Bericht 2010–04, Institut für Informatik der Universität Augsburg (2010)
6. Wirsching, G., Kölbl, C., Huber, M.: Zur Logik von Bestenlisten in der Dialogmodellierung. In: Kröger, B.J., Birkholz, P. (Hrsg.) Elektronische Sprachsignalverarbeitung 2011, Tagungsband der 22. Konferenz, Aachen, 28. bis 30. Studientexte zur Sprachkommunikation, vol. 61, pp. S.309–S.316 (September 2011)
7. Wirsching, G.: Semirings Modeling Confidence and Uncertainty in Speech Recognition, Preprint Mathematik, KU Eichstätt-Ingolstadt (2011), http://edoc.ku-eichstaett.de/6083/
8. Wirsching, G., Kölbl, C.: Language Modeling with Utterance-Meaning Pairs. Technischer Bericht 2011–12, Institut für Informatik der Universität Augsburg (2011)

Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström

Department of Speech, Music, and Hearing, KTH Royal Institute of Technology,
Lindstedtsvägen 24,
10044SE Stockholm, Sweden
{sameram,beskow,skantze,bjorn}@speech.kth.se
http://www.speech.kth.se

Abstract. In this chapter, we first present a summary of findings from two previous studies on the limitations of using flat displays with embodied conversational agents (ECAs) in the contexts of face-to-face human-agent interaction. We then motivate the need for a three dimensional display of faces to guarantee accurate delivery of gaze and directional movements and present *Furhat*, a novel, simple, highly effective, and human-like back-projected robot head that utilizes computer animation to deliver facial movements, and is equipped with a pan-tilt neck. After presenting a detailed summary on why and how *Furhat* was built, we discuss the advantages of using optically projected animated agents for interaction. We discuss using such agents in terms of situatedness, environment, context awareness, and social, human-like face-to-face interaction with robots where subtle nonverbal and social facial signals can be communicated. At the end of the chapter, we present a recent application of *Furhat* as a multimodal multiparty interaction system that was presented at the London Science Museum as part of a robot festival,. We conclude the paper by discussing future developments, applications and opportunities of this technology.

Keywords: Facial Animation, Talking Heads, Robot Heads, Gaze, Mona Lisa Effect, Avatar, Dialogue System, Situated Interaction, Back Projection, Gaze Perception, Furhat, Multimodal Interaction, Multiparty Interaction.

1 Introduction

There has always been an urge in humans to give machines an anthropomorphic appearance and behavior. This urge, perhaps, comes from the human interest to understand and recreate themselves, since humans can be considered (or at least appear to be) the most intelligent and complex animations of life.

This orientation of giving machines a human body and face has been clear since the beginning of works on robotics. For example, the word “robot” was introduced to the public by the Czech interwar writer Karel Čapek in his play R.U.R. (Rossum's Universal Robots), published in 1920. The play begins in a factory that makes artificial people called robots, though they are closer to the modern ideas of androids, creatures that can be mistaken for humans [1].

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 114–130, 2012.

The Holy Grail in the quest for building human-like robots, however, has been the human face. Simulating the appearance and dynamics of the human face has been shown to be an intensely complex matter. The human face, with its subtle and minute movements, carries an incredible amount of information that is designed to be read and interpreted by others. For instance, the human lips carry significant information about speech and intonation [2] [3], the eyes are a mirror to the mind, affect and attention ([4] [5]). The combination of these components provides the human with the possibility to communicate emotions as well as interests. However, it also provides information about more physical parameters such as age and gender, ([6]).

The efforts for building natural anthropomorphic faces has mainly taken two different tracks; one building of physical, mechanical heads that simulate the structure and appearance of a human face; and the other one has been focusing on building three dimensional digital animated computer models. Figure 1 illustrates examples for both tracks.

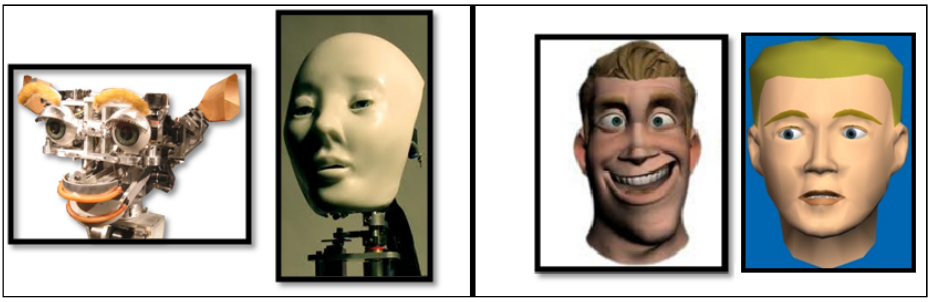
Building computer simulations of the human face has indeed been a challenging task, but recently making impressive progress. This is mainly due to its major applications in the gaming and moving-picture industries, those being the driving forces behind much of the progress. These models have also been intensively used as a research tool to better understand the functionality of the human face, taking advantage of the flexibility and easy manipulation of these models. An important advantage of these computer models is that they can be replicated at no cost, providing different branches of research and industry with very good accessibility.

Unfortunately, this advancement has not been paralleled in robotics in general: The easy control of computer models is not easily mapped onto control of muscular and mechatronic movements of servos implemented in robotic heads [7], introducing huge limitations in human-looking robotic faces to exhibit smooth and human-like movement, and hence introducing inconsistencies between how the robot looks and how it behaves (usually referred to as the uncanny valley [8]). The other limitation of building human-like robotic faces is their expensive manufacturing and replication. At the moment, there are only a handful of human-like robots, which is making them exclusive and inaccessible to both the research community and the public.

Some trials have been carried out to bridge this gap between software animation (virtual agents) and physical robots. One solution has been to use a computer screen as a robot head [9], with a virtual agent embedded into it. This approach offers a face with natural looks and dynamics while preserving a physical robot body. However, it naturally suffers several limitations and problems that come with using a flat display as an alternative to a three dimensional physical head, such as that, (aside from large aesthetic inconsistencies), flat displays are not three dimensional and suffer from lacking absolute direction of what is presented into them in relation to where the screen is placed (more detailed discussion in Section 2).

In this chapter, we are presenting a highly natural and effective hybrid solution for using animated agents for robotic heads. We are building on two previous studies that demonstrate the limitations of flat screens in delivering accurate direction of gaze, and hence limit the capabilities of animated agents to carry out situated, multiparty interaction. After that, we present *Furhat*, a three dimensional back-projected robot

head that utilizes a computer animated face. We describe the details on how *Furhat* was built and what advantages it offers over, both in-screen animated agents, and mechanical robotic heads. After that we discuss possible applications of using *Furhat* for multimodal, multiparty human-machine interaction, and demonstrate a system for a three-party dialogue with *Furhat* which has recently been showcased at the London Science Museum as part of a European robot festival.



(a) (b)

Fig. 1. (a) Two examples of physical robot heads. (b) Two examples of computer animated facial models.

2 Animated Agent and Mechanical Robots

As discussed earlier, interactive agents that are made to look and act as humans can come in two instantiations. First as virtual characters (where the body and face of the agent is a computer software), or second, as physical robots.

One may think of robots as situated physical agents: At the time of interaction, the agent and the human are co-present spatially and temporally, which ultimately simulates the human-human communication setup. However, virtual agents are computer software that are, clearly, not co-present spatially with the interactive partner (the human) in the same space, but can be thought of as living in a virtual space. Many approaches have been tried to optimally bridge these two physical and virtual worlds, and bring the human and the virtual agent into the same world. Those being virtual reality interfaces (Figure 2 left), and holographic projections (Figure 2 right).

In virtual reality, pragmatically, the human is transferred into the three dimensional virtual world, while in holographic projection, the virtual three dimensional world is transferred into our own reality, and hence, both co-exist spatially with the human interlocutor.

These two solutions are highly complex, exclusive and expensive, and are seldom used as a user interface with virtual characters. However, the predominant solution to bridging the virtual and the real worlds has been via projections onto flat displays (such as flat screens, wall projections, etc.); an example is shown in the middle of Figure 2. The flat display functions as a window between the world the human interlocutor is situated in, and the virtual world of the virtual character [10].

(a) (b) (c)

Fig. 2. (a) An example of a person wearing virtual reality (VR) glasses, so to be immersed in a virtual world. (b) An example of a virtual character that is presented via a flat display, offering a bridge into the physical and virtual realities. (c) An example of a holographic display of a person, to bring the virtual character into the physical space.

It is known that the perception of three-dimensional objects that are displayed on two-dimensional surfaces is guided by, what is commonly referred to as the Mona Lisa effect [11]. This means that the orientation of the three-dimensional objects in relation to the observer will be perceived as constant, no matter where the observer is standing in the room or in relation to the display. For example, if the portrait of a face is gazing forward, mutual gaze will be established between the portrait and the observer, and this mutual gaze will hold no matter where the observer is standing. Accordingly, if the portrayed face is gazing to the right, everyone in the room will perceive the face as looking to their left. Thus, either all observers will establish mutual gaze with the portrait or none of them will. This implies that no exclusive eye-contact between the portrait and only one of the observers is possible. This principle, of course, extends to all objects viewed on 2D surfaces, such as pointing hands or arrows.

This effect can be seen as the cost of bridging the two different, virtual and real, worlds, to allow for direct visual interaction between humans and animated agents. This effect, clearly, has important implications on the design of interactive systems, such as embodied conversation agents, that are able to engage in situated interaction, as in pointing to objects in the environment of the interaction partner, or looking at one exclusive observer in a crowd.

In the following two sections we will present the results from two previous studies showing the limitations of the Mona Lisa effect on interaction, and presenting an approach on extending the use of animated faces from the flat screen onto physical three dimensional head models (and so building a physical situated robotic head). These two studies represent a proof of concept of this approach to overcome the limitations of flat displays of animated faces.

3 Background Study 1: Perception of Gaze

Since the Mona Lisa gaze effect is introduced by 2D projection surfaces, we suggested an alternative to 2D projection surfaces, by which the Mona Lisa gaze effect would be avoided. Our approach in this experiment was to use a 3D physical, static model of a human head (as seen in Figure 3). In order to compare this model with a traditional 2D projection surface, we designed an experimental paradigm that tests for mutual gaze as well as for gaze direction in the physical space of the viewer. The method is used to test the differences in accuracies in predicting gaze direction from a face that is presented through a 2D surface and the 3D projected surface.

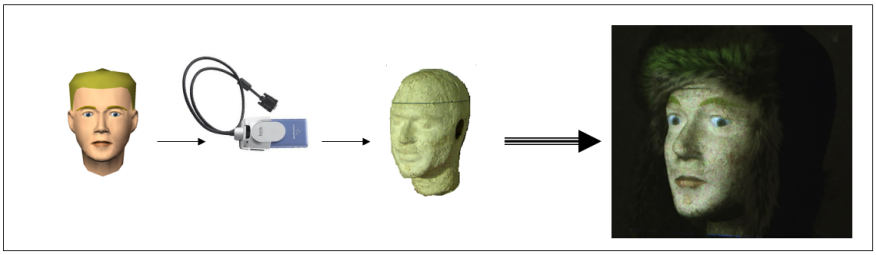


Fig. 3. An earlier approach for front projecting an animated face onto a physical head model using a micro laser projector

The technique of manipulating static objects with light is commonly referred to as the *Shader Lamps* technique [12] [13]. This technique is used to change the physical appearance of still objects by illuminating them using projections of static or animated textures, or video streams.

In the perception experiment in [14], five subjects were simultaneously seated around an animated agent, which shifted its gaze in different directions (see Figure 4). After each shift, each subject reported who the animated agent was looking at. Two different versions of the same head were used, one projected on a 2D surface, and one projected on a 3D static head-model (see Figure 5). The results showed a very clear Mona Lisa effect in the 2D setting, where all subjects perceived a mutual gaze with the head at the same time for frontal and near frontal gaze angles.

While the head was not looking frontal, none of the subjects perceived mutual gaze with the head. In the 3D setting, the Mona Lisa effect was completely eliminated and the agent was able to establish mutual and exclusive gaze with any of the subjects. The subjects achieved a very high agreement rate on guessing on which subject the gaze of the agent was directed at for all the different gaze shifts.

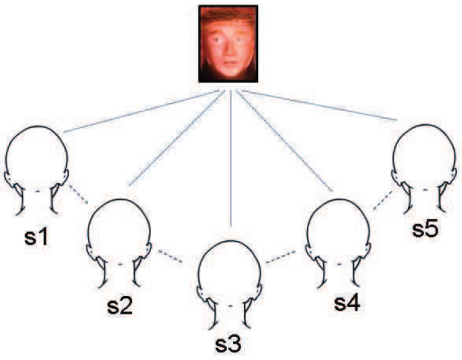


Fig. 4. Schematic setup and placement of the subject and stimuli point

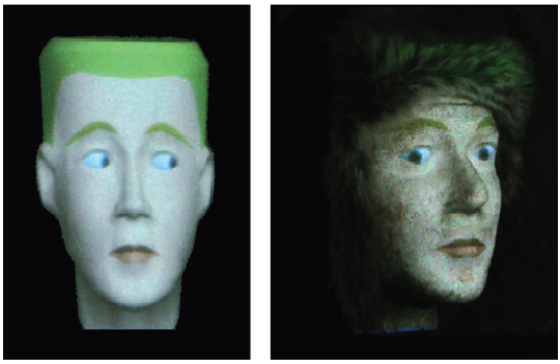


Fig. 5. A snapshot of the animated agent displayed on a 2D white board (left), and on a 3D head model (right)

This study provides important insights and proves the principal directional properties of gaze through a 2D display surface. The study also shows that using the simple approach of optically projecting the same face model onto a 3D physical head model would eliminate that effect. However, the study does not show whether this effect will hold during interaction, or whether people are able to cognitively compensate for the effect, and correctly infer the *intended* direction of gaze.

4 Background Study 2: Interactional Effects of Gaze

In order to explore the interactional effects of gaze in a multi-party conversational setting, a similar experiment was carried out, but with spoken interaction between the

head and the participants [15]. Unlike the previous perception experiment, which focused on the *perceived* gaze, this experiment investigated how gaze may affect the turn-taking *behavior* of the subjects, depending on the use of 2D or 3D displays.

Two sets of five subjects were asked to take part in the experiment. In each session, the five subjects were seated at fixed positions at an equal distance from each other and from an animated agent (just as in the previous experiment, see Figure 4). The agent addressed the subjects by directing its gaze in their direction. Two versions of the agent were used, one projected on a 3D head model and one projected on a flat surface, as shown in Figure 5. The conversational behavior of the animated agent was controlled using a Wizard-of-Oz setup. For each new question posed by the agent, the gaze was randomly shifted to a new subject. The subjects were given the task of watching a video from a camera navigating around the city of Stockholm, after which the animated agent asked them to describe the route they had just seen. After each video was finished, the animated agent started to ask the subjects about directions on how to reach the landmark the video ended with, starting from the point of view the video started with. Each set of subjects did four dialogs in both the 2D and the 3D condition (i.e. a total of eight videos).

To measure the efficiency of the gaze control, a confusion matrix was calculated between the intended gaze target and the actual turn-taker. The accuracy for targeting the intended subject in the 2D condition was 53% and 84% for the 3D condition. The mean response time was also calculated for each condition, i.e. the time between the gaze shift of the question and the time takes for one of the subjects to answer, which showed a significant difference in response time between the two conditions: 1.86 seconds for the 2D condition vs. 1.38 seconds in the 3D condition.

The results show that the use of gaze for turn-taking control on 2D displays is limited due to the Mona Lisa effect. The accuracy of 50% is probably too low in settings where many users are involved. By using a 3D projection, this problem can be avoided to a large extent. However, the accuracy for the 2D condition was higher than what was reported in the previous experiment. A likely explanation for this is that the subjects in this task may to some extent compensate for the Mona Lisa effect – even if they do not "feel" like the agent is looking at them, they may learn to associate the agent's gaze with the intended target subject. This comes at a cost, however, which is indicated by the longer mean response time. The longer response time might be due to the greater cognitive effort required making this inference, but also to the general uncertainty among the subjects about who is supposed to answer.

The subjects were also asked to fill out a questionnaire after the interactions, in which they compared the two versions of the head along three dimensions, as shown in Figure 6. As the figure shows, the 3D version was clearly preferred, perceived as more natural, and judged as less confusing when it comes to knowing whose turn it was to speak.

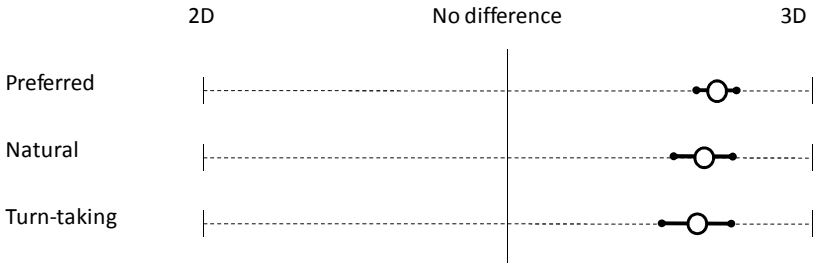


Fig. 6. The subjective assessment of the 2D and 3D versions of the talking head, showing mean and standard errors

5 The Furhat Robot Head

As shown in the previous studies and discussions above, the paradigm of using a physical head model as a projection surface for animated computer models, would not only bring the face outside of the traditional two-dimensional screen, but will also eliminate the Mona Lisa effect and allow for multiparty interaction. From the study above in Section 4, it also appears that people perceive the projected face as significantly more natural than the face shown inside the screen. In addition to that, using the animated computer model as an alternative to a physical robot head solves major difficulties for building naturally looking and moving robot faces, since the technology behind facial animation has reached impressive advancements, and the control of these faces is highly simple and flexible. (Refer to [16] for a short review on the benefits of this approach).

Building on these encouraging findings, we have started building a natural and human-like robotic head that is based on the principle of optically projected computer models. A main modification was applied to the previous approach; that is to back-project the face onto the mask, so that the projector is hidden behind the mask. This means that if the mask is placed onto a robotic neck, the mask and the projector will be attached together and the projected image will not be displaced.

To build the head, several factors had to be taken into account. For example, micro projectors have a small projection angle, and hence if the projector is placed too close to the mask, the projected image will not be big enough to cover the entire projection area of the mask. Another factor was to use a material that will diffuse the light over the mask so that the light projected on the mask will be equally illuminated. One last important factor that had to be taken into account is to be able to acquire a mask model that would exactly fit the design of the projected face, so that no calibration and transformations of the model will be needed, and subtle facial areas, like the eyes, will naturally fit the area of the eyes on the mask.

Figure 7 shows a flow chart of the process of how the back-projected head is built. We call the head *Furhat*, as it got a fur hat that covers the top and the sides of the mask. Following is a detailed description on how *Furhat* has been built, so that it would provide more insights into the properties of the head, and comes as a guide for others to replicate it.

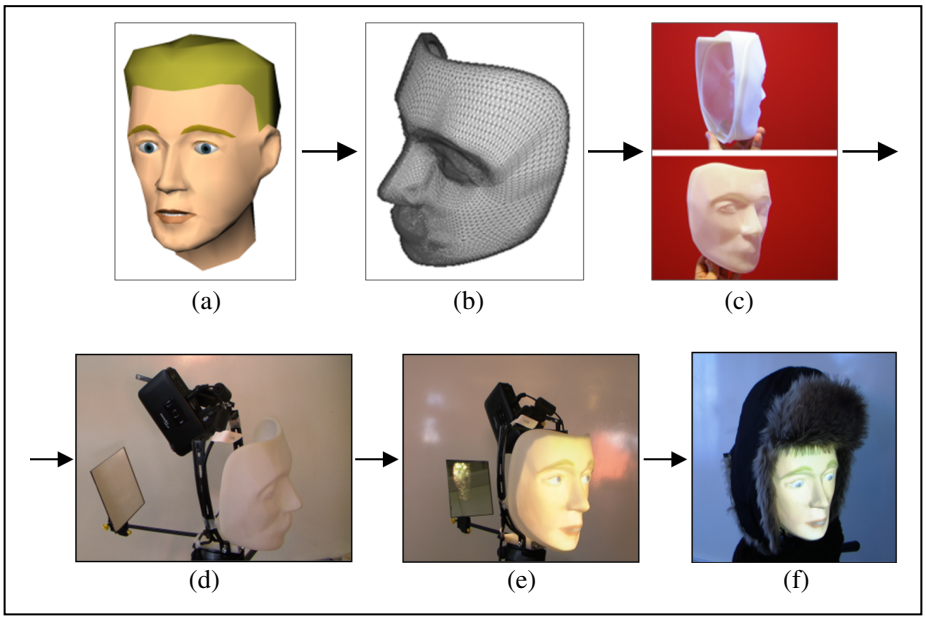


Fig. 7. A chart showing the process for building Furhat, the back projected robotic head

Building Furhat

In the following section we provide a chronological list of the main steps taken to build the robot head:

1- Using an animated face model: The 3D animated face model that is used for this study is detailed in [17]. An animated face model is used due to several reasons: The lips of the face model can be automatically synchronized with the speech signal the system is producing; this is done by using a transcription of the speech utterances to be produced. The lip synchronization system utilized in the face model has proven to enhance speech intelligibility over listening only to the audio signal [18]. This face model also offers flexible control of gestures and facial movements (gaze movements, eyebrows movements, etc.). Gestures played using this face model have also been shown to deliver the communicative functions they are designed for (eyebrows raise to signal questioning, doubt, or surprise [19]); these gestures have also been shown to enhance speech intelligibility [20].
From this evidence, it is clear that this face model can deliver highly accurate and natural movements and would be suitable as a choice for *Furhat*'s face.

2- Printing the 3D mask: The main step is to establish a translucent mask that would allow the back projected light to be clearly visible when looked at from the front. The other important factor is to establish a mask that fits in its shape,

the face model that would be projected on top of it (mentioned in the previous point). To establish this, a 3D copy of the exact face 3D model was printed using a 3D printer, with an equal overall thickness of 1mm. After sample testing, this thickness proved optimal to allow just enough light to be visible on the mask. Figure 7a shows the original 3D computer model of the face. Figure 7b shows the 3D design of the mask acquired by modifying the original face model, and making it suitable for 3D printing. Figure 7c shows the mask after printing. The dimensions of the printed mask were made to resemble the size of an average human head (width 16cm, height 22cm, depth 13cm).

3- Allowing the mask to equally diffuse color: A main problem of back projecting light on translucent objects is that the light-source will be visible (glowing) when looked at from the front. This was an obvious problem when the printed face was used with a micro projector. To solve this problem, a back-projection paint, which is used to create back-projection screens, was used (goo systems Global[1]). This spray paint is used specifically to allow the cured surface to diffuse the light[1] equally over its surface, and hence diminish the problem of unbalanced optical illumination over the mask. Figure 7e shows the back-projected face after applying the back-projection paint on the plastic mask.

Fig. 8. A front and back view of the mask and the rig of Furhat

4- Rigging the mask with a micro projector: When the mask was tested and proved ready to use as the back projection mask for the head, the mask then was rigged with a micro-projector that was placed on top of the mask, the projector then projects light onto a mirror that reflects back the face onto the mask. This approach allows for more distance between the projector and the mask, which in turn, allows for the projected image to be in focus and to fit the entire mask.

[1] http://www.goosystemsglobal.com/

Figure 7d shows how the head is rigged with a projector and a mirror. Figure 8 shows a front and back view of the head when the mask is rigged with the projector and a mirror, showing how the projected face fits exactly the 3D plastic mask (it is important to note here that the solution of using a mirror is probably replaceable by other alternatives such as using a fish-eye lens that widens the projection area of the projector). After the mask was rigged, the head was covered using a fur hat. The fur hat covers the projector and the rig, and hence gives a stronger focus on the facial appearance of *Furhat*. Figure 9 shows *Furhat* with and without its head cover.

Fig. 9. Snapshots of Furhat with and without the head cover (the fur hat)

5- Giving *Furhat* a neck:

Direction of attention may of course not only be achieved with the eyes, but also by moving the head, using a neck. A neck allows the robot head to use either eye movement, head pose, or both, to direct the attention, but also to do gestures such as nodding. Depending on which behaviors need to be modeled, different degrees of freedom (DOF) may be necessary. To direct the gaze in any direction (if the eyes are centered), 2 DOF are obviously necessary, but in order to perform a wider range of gestures, more DOF may be needed. An example of a very flexible robot neck is presented in [21], where 3 DOF are used: lower and upper pitch (tilting up and down), yaw (panning side to side) and rolling (tilting side to side). Lower pitch is centered where the neck meets the shoulders, and high pitch is centered where the neck is attached to the head.

For *Furhat*, we are currently using a pan-tilt unit. The unit has a no-load speed of 0.162 sec/60° and a holding torque of 64 kg·cm. It has 2 DOF: pitch and yaw, which allows *Furhat* to direct the head in any direction, but also to do simple gestures such as nodding.

6 Example Application

The development of *Furhat* is part of a European project called IURO (Interactive Urban Robot)[2]. As part of this project, we were invited to the EUNIC RobotVille Festival at the London Science Museum, December 1st – December 4th, 2011. The purpose of the IURO project is to develop robots that can obtain missing information from humans through multi-party dialogue. The central test-case will be an autonomous robot that can navigate in an urban environment by asking humans for directions. For the exhibition, we wanted to explore a similar problem, but to suit the setting we instead gave *Furhat* the task of asking the visitors about their beliefs of the future of robots, with the possibility of talking to two visitors at the same time and shifting attention between them.

In lab setups, we have been using Microsoft Kinect[3], which includes a depth camera for visual tracking of people approaching *Furhat* and an array microphone for speech recognition. However, due to the crowded and noisy environment in the museum, we chose to use handheld close-range microphones and ultrasound proximity. For speech recognition, the Microsoft Speech API was used. For speech synthesis, we used the CereVoice William TTS from CereProc[4]. CereVoice reports the timing of the phonemes in the synthesized utterance, which was used for synchronization of the lip movements in the facial animation. It also contains a number of verbal gestures that were used to give *Furhat* a more human-like appearance, such as grunts, laughter and yawning.

To control *Furhat*'s behavior, we used an event-driven system implemented in Java, inspired by Harel state-charts [22] and the UML modeling language. This allowed the system to react to external sensory input (speech, proximity data) as well as self-monitoring data, and produce actions such as speech, facial gestures and head movements. The layered structure of the state-chart paradigm allows the dialogue designer to define a hierarchy of dialogue states, and the sensory-action paring that is associated with these states. For the exhibition scenario, the dialogue contained two major states reflecting different initiatives: one where *Furhat* had the initiative and asked questions to the visitors (i.e., "when do you think robots will beat humans in football?") and one where the visitors asked questions to *Furhat* (i.e., "where do you come from?"). In the former case, *Furhat* continued the dialogue (i.e., "why do you think so?"), even though he often understood very little of the actual answers, occasionally extracting important keywords.

With nobody close to the proximity sensors, *Furhat* was in an "idle" mode, looking down. As soon as somebody approached a proximity sensor, he looked up and initiated a dialogue with "Could you perhaps help me?". The multi-party setting allowed us to explore the use of head-pose and gaze during the dialogue:

[2] http://www.iuro-project.eu/
[3] http://kinectforwindows.org/
[4] http://www.cereproc.com/

- With two people standing in front of him, *Furhat* was able to switch interlocutor using first a rapid gaze movement and then head movement. Often *Furhat* used this possibility to move the dialogue forward, by switching interlocutor and asking a follow-up, such as "do you agree on that?"
- *Furhat* could either ask a specific interlocutor, or direct the head between the interlocutors and pose an open question, moving the gaze back and forth between the interlocutors. By comparing the audio-level and timing of the audio input from the two microphones, *Furhat* could then choose who to attend and follow-up on.
- If *Furhat* asked a question specifically to one of the interlocutors, and the other person answered, he quickly used gaze to turn to this person saying "could you just wait a second", then shifted the gaze back and continued the dialogue.

To exploit the possibilities of facial gestures that the back-projection technique allows, certain sensory events were mapped to gesture actions in the state chart. For example, when the speech recognizer detected a start of speech, the eyebrows were raised to signal that *Furhat* was paying attention.

Fig. 10. Furhat at the London Science Museum. The monitor shows the results of the visitors' answers to Furhat's questions. The two podiums with microphones and proximity sensors can also be seen.

In total, 7949 people visited the exhibition during the course of 4 days. The system proved to be very stable during the whole period. Apart from the video data, we recorded 8 hours of speech from the visitors. We also let the visitors fill out a questionnaire about their experience after the interaction. We have not yet analyzed

the data, but it was apparent that many visitors liked the interaction and continued to answer *Furhat*'s questions although he actually understood very little of their answers. The visitors also seemed to understand *Furhat*'s attentive behavior and act accordingly. Videos from the exhibition can be seen at www.speech.kth.se/furhat.

7 Discussions

One major motivation behind this work is to build a robot head that can use state-of-the-art facial animation to communicate and interact with humans. These include natural and smooth lip movements, control of perceivable eye and gaze movements.

To make a robot head that is able to capitalize on social signals, its head should be able to generate such signals to highly perceivable accuracy. The first step towards reaching this goal was to use animated talking agents. However, since the robot is supposed to be able to engage in interactive multimodal dialogue with multiple people, the simple solution of using a computer screen as an interface with an animated agent projected onto it became disadvantageous. This is due to the fact that the 2D screen has no direction, and suffers from the Mona Lisa gaze effect (amongst other effects). This effect makes it impossible to establish, for example, exclusive eye-contact with one person out of many.

The solution to reach these goals, while avoiding the hindering effects of flat displays, is *Furhat*, a hybrid solution that can be thought of as bringing the animated face out of the screen and into the real-physical world.

Fig. 11. Examples showing different instantiations of the colors of Furhat's facial features

Clearly, the benefits of using an animated agent as a robot head employing optical projection meets the goal of bringing the smooth and accurate animation of 3D computer models into a robot head. But there are more advantages. The flexibility of using a computer model allows for fast and online control of the face depending, for example, on context. *Furhat* for example can change its facial design on the fly since the colors and shape of its different facial parts is just a software animation (this

manipulation is however limited by the design of the mask). Figure 11 shows examples of different facial colors of *Furhat*.

These and other parameters can be controlled depending on context and the environment, for example, *Furhat* can have a different facial design depending on the cultural background, or the age of the interlocutor. It can change its color contrasts depending on the surrounding light.

One expressive and environment-sensitive part of the face that can be controlled in this setup is the eyes. The pupil size for instance, can correspond to the amount of light in the surroundings [23], and can also reflect functions such as affect and interest.

Another context-aware property of the eyes is the corneal reflection. This is when the image of the environment is reflected on the cornea. This phenomenon has been shown to provide significant amount of information about the environment where the interaction is taking place [24].

These features can be easily implemented in *Furhat* on the software side by controlling the size and textures of the eyes and hence the projected image will more accurately reflect the situated context *Furhat* is interacting in.

Other benefits of *Furhat* to be used as a robotic head are its low weight, low maintenance demands, low noise level (only the noise coming from the neck), and its low energy consumption.

8 Conclusions

In this chapter, we have presented *Furhat*, an example of a paradigm for building robot heads using software based animated faces. Based on experimental evidence, this paradigm makes animated faces look more natural and human-like since it brings them out of the screen and onto a human-head-shaped three-dimensional physical object. This, not only makes animated faces look more natural in interactions, but also solves problems that arise when visualizing them onto flat displays. Such potential problems are achieving accurate multiparty interaction using gaze and head direction (since flat displays lack the enforcement of direction).

Looking at what *Furhat* has to offer to robotic heads, the advantages of using software design and animation instead of hardware (physical-mechanical) design and animation are numerous. Robot heads lack the ability to move their facial parts smoothly and accurately enough to simulate human facial movements (eye movements, blinking, eyebrows movement, and specially lip movements), let alone looking like human ones.

Furhat, on the other hand, uses an animated face that can move its facial parts online, in real-time, and to a large degree like humans do. In addition to movement, the design of the face is very flexible. The design of robot heads typically cannot change after manufacturing the head (the color and design of the lips and eyebrows, the color of the eyes, the size of the iris…), *Furhat*'s colors and design, on the other hand, can easily change. This is achieved by using the animated face model it utilizes as its face, while still using the same face mask and hardware, and hence no mechanical or hardware cost is associated with this functionality.

After we presented *Furhat* and how it was built in this paper, allowing for others to possibly replicate the process, we have presented a sample application that uses *Furhat* for multiparty interaction with human, which was presented at the London Science Museum for 4 days and received around 8000 visitors.

We would like to use *Furhat* not only as a natural interactive robot head, but also as a research framework which allows for studying human-human (one can think of *Furhat* as a tele-presence device) and human-robot interaction in single and multiparty setups and in turn-taking and dialogue management techniques using face and neck movements, to count a few.

Acknowledgments. This work has been done at the Department for Speech, Music and Hearing, and funded by the EU project IURO (Interactive Urban Robot) No. 248314. The authors would like to thank Simon Alexanderson for designing the 3D mask model for printing, and to thank Jens Edlund, Joakim Gustafson and Preben Wik for their interest and inspiring discussions.

References

1. Dominik, Z.: Who did actually invent the word robotand what does it mean? The Karel Čapek website, `http://capek.misto.cz/english/robot.html` (retrieved December 10, 2011)
2. Summerfield, Q.: Lipreading and audio-visual speech perception. Philosophical Transactions: Biological Sciences 335(1273), 71–78 (1992)
3. Al Moubayed, S., Beskow, J.: Effects of Visual Prominence Cues on Speech Intelligibility. In: Proceedings of Auditory-Visual Speech Processing, AVSP 2009, Norwich, England (2009)
4. Argyle, M., Cook, M.: Gaze and mutual gaze. Cambridge University Press (1976)
5. Kleinke, C.L.: Gaze and eye contact: a research review. Psychological Bulletin 100, 78–100 (1986)
6. Ekman, P., Friesen, W.V.: Unmasking the face: A guide to recognizing emotions from facial clues. Malor Books (2003) ISBN: 978-1883536367
7. Shinozawa, K., Naya, F., Yamato, J., Kogure, K.: Differences in effect of robot and screen agent recommendations on human decision-making. International Journal of Human Computer Studies 62(2), 267–279 (2005)
8. Mori, M.: Bukimi no tani.:The uncanny valley (K. F. MacDorman & T. Minato, Trans.). Energy 7(4), 33–35 (1970) (Originally in Japanese)
9. Gockley, R., Simmons, J., Wang, D., Busquets, C., DiSalvo, K., Caffrey, S., Rosenthal, J., Mink, S., Thomas, W., Adams, T., Lauducci, M., Bugajska, D., Perzanowski, Schultz, A.: Grace and George: Social Robots at AAAI. In: Proceedings of AAAI 2004, Mobile Robot Competition Workshop, pp. 15–20. AAAI Press (2004)
10. Edlund, J., Al Moubayed, S., Beskow, J.: The Mona Lisa Gaze Effect as an Objective Metric for Perceived Cospatiality. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS (LNAI), vol. 6895, pp. 439–440. Springer, Heidelberg (2011)
11. Todorovi, D.: Geometrical basis of perception of gaze direction. Vision Research 45(21), 3549–3562 (2006)

12. Raskar, R., Welch, G., Low, K.-L., Bandyopadhyay, D.: Shader lamps: animating real objects with image-based illumination. In: Proc. of the 12th Eurographics Workshop on Rendering Techniques, pp. 89–102 (2001)
13. Lincoln, P., Welch, G., Nashel, A., Ilie, A., State, A., Fuchs, H.: Animatronic shader lamps avatars. In: Proc. of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2009). IEEE Computer Society, Washington, DC (2009)
14. Al Moubayed, S., Edlund, J., Beskow, J.: Taming Mona Lisa: Communicating gaze faithfully in 2D and 3D facial projections. ACM Trans. Interact. Intell. Syst. 1(2), Article 11, 25 pages (2012)
15. Al Moubayed, S., Skantze, G.: Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. In: Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP, Florence, Italy (2011)
16. Al Moubayed, S., Beskow, J., Edlund, J., Granström, B., House, D.: Animated Faces for Robotic Heads: Gaze and Beyond. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 19–35. Springer, Heidelberg (2011)
17. Beskow, J.: Talking heads - Models and applications for multimodal speech synthesis. Doctoral dissertation, KTH (2003)
18. Beskow, J.: Animation of talking agents. In: Benoit, C., Campbel, R. (eds.) Proc of ESCA Workshop on Audio-Visual Speech Processing, Rhodes, Greece, pp. 149–152 (1997)
19. Granström, B., House, D.: Modeling and evaluating verbal and non-verbal communication in talking animated interface agents. In: Dybkjaer, l., Hemsen, H., Minker, W. (eds.) Evaluation of Text and Speech Systems, pp. 65–98. Springer (2007)
20. Al Moubayed, S., Beskow, J., Granström, B.: Auditory-Visual Prominence: From Intelligibility to Behavior. Journal on Multimodal User Interfaces 3(4), 299–311 (2010)
21. Brouwer, D.M., Bennik, J., Leideman, J., Soemers, H.M.J.R., Stramigioli, S.: Mechatronic Design of a Fast and Long Range 4 Degrees of Freedom Humanoid Neck. In: Proceedings of ICRA, Kobe, Japan, ThB8.2, pp. 574–579 (2009)
22. Harel, D.: Statecharts: A visual formalism for complex systems. Science of Computer Programming 8(3), 231–274 (1987)
23. Blackwell, R.D., Hensel, J.S., Sternthal, B.: Pupil dilation: What does it measure? Journal of Advertising Research 10, 15–18 (1970)
24. Nishino, K., Nayar, S.K.: Corneal Imaging System: Environment from Eyes. Int. J. Comput. Vision 70(1), 23–40 (2006), doi:10.1007/s11263-006-6274-9

VISION as a Support to Cognitive Behavioural Systems*

Luca Berardinelli[1], Dajana Cassioli[1], Antinisca Di Marco[1],
Anna Esposito[2], Maria Teresa Riviello[2], and Catia Trubiani[1]

[1] University of L'Aquila, Dipartimento di Informatica, Italy
[2] Second University of Naples, Department of Psychology and IIASS, Italy
{luca.berardinelli,dajana.cassioli,antinisca.dimarco,
catia.trubiani}@univaq.it,
{anna.esposito, mariateresa.riviello}@unina2.it

Abstract. Cognitive behavioral systems would definitely benefit from a supporting technology able to automatically recognize the context where humans operate, their gestures and even facial expressions. Such capability poses challenges for many researchers in various fields because the ultimate goal is to transfer to machines the human capability of representing and reasoning on the environment and its elements. The automation can be achieved through a supporting infrastructure able to capture a huge amount of information from the environment, much more than humans do, and sending it to a processing unit able to build a representation of the context that would catch all elements necessary to interpret the specific environment.

The goal of this paper is to present the VISION infrastructure and how it can support cognitive systems. Indeed, VISION is a software/hardware infrastructure that overcomes the limitations of current technology for Wireless Sensor Networks (WSNs) providing broadband wireless links for 3D video streaming with very high reliability, obtained by an innovative reconfigurable context and resource aware middleware for WSNs. We show VISION at work on the communicative impaired children scenario.

Keywords: Wireless Sensor Networks, Reconfigurable Infrastructure, Broadband Communication Channel, Cognitive Systems.

1 Introduction

Human perception is a constructive process that provides the individual with all information that is sufficient to formulate a good representation of the context, to make him conscious of its own presence and of presence of other people, to recognize multi modal interactional information such as facial expressions and body gestures.

The available inputs from the context are countless, but the human perception mechanisms are able to operate an intelligent selection before the relevant information are processed and correlated. For instance, the stereoscopic vision, enhanced by fast movements of eyes, provides the necessary redundancy of visual information, which results, once appropriately processed, in the 3D reconstruction of surroundings.

* This work has been supported by the EU-funded VISION ERC project (ERC-240555).

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 131–143, 2012.

When vision is correlated to a number of other sensorial information, like audio, temperature, light, body signal parameters, etc., the surrounding context is completely figured out by a human being.

Providing a sensing machine with this powerful intelligence is the ambitious aim of many research fields. The ultimate goal would be transferring to machines the human capability of representing the context in order to develop several types of applications able to reason on it. The automatic recognition of a given status of the context, of a human being, and of the interactional information exchange, will push up current technology in security for surveillance applications, in environmental monitoring, in remote and local human-machine interaction, or in the human-human interaction even in case of impaired people.

The first stage of the process for the intelligent perception of reality is the acquisition of environmental data (e.g., audio, video, temperature, human body signal parameters, etc). Of course, machines cannot select a priori the required elements to perceive the context, their task is to capture a huge amount of information from the environment. Such information must be managed by a processing unit able to build a representation of the context that would catch all elements necessary for the specific application: a full detailed 3D reconstruction is suitable for virtual reality applications, whereas a focus on a shaking hand, neglecting all other details, is sufficient for recognizing the gesture.

Recently, growing interest is focusing on cognitive systems [6,8] that are systems embedding and using psychological data, similar to the ways humans think and process information. They engage the functions of human cognition and increase one's cognitive capabilities helping, for example, people with difficulties on processing of information.

Cognitive systems should follow the users remaining transparent to them, in other words they must be pervasive and ubiquitous. Moreover, they need to capture and elaborate context information for the users. These aims can be reached by relying on wireless technologies able to transparently sense the context in terms of audio, video and typical sensor data.

This paper aims at showing how cognitive systems can be built on top of VISION [1] that is an infrastructure for intelligent ubiquitous sensing services.

Ubiquitous sensing services are nowadays provided by very simple devices, with small storage capacities and low performance processing units. These communicate using low-data rate channels that prevent the video streaming service. VISION overcomes the limitations of current technology for Wireless Sensor Networks (WSNs), because it is suitable to capture and carry a lot of information about the context.

VISION supports a set of sensing services, customizable over the specific application, but privileging the real-time 3D video sensing. In particular, VISION proposes to re-design WSNs by employing powerful devices equipped by a very broad-band radio communication system and optimizing the use of resources and the operation mode of these devices by means an innovative middleware supporting context-aware techniques and its adaptation to the context. For context we intend both the internal status of the devices and the status of the surroundings. The optimization of the internal status of every device guarantee high reliability of the WSN, increased lifetime and energy saving. The optimization of the device behavior with respect to the instantaneous external

context assures that the best system's configuration and tailored services are used for the specific application scenario.

In this paper we focus on the scenario of communicative impaired children and show how an application can be designed relying on VISION services in order to support and improve the learning processes of those children. Thanks to the new wireless technology and the innovative reconfigurable middleware, the impaired children and their caregiver can be provided with a transparent and resilient system supporting the teaching (caregiver) and the learning (impaired children) processes. Note that the VISION infrastructure is ongoing work, however we aim to provide a demo to show the applicability of the scenario we present in this paper.

The paper proceeds by describing the VISION infrastructure (Section 2), specifying the cognitive application designed to support both care givers and children in learning process of impaired children (Section 3) and how this application can be designed with the support of VISION infrastructure (Section 4). Finally, Section 5 reports on the related work present in the literature and Section 6 concludes the paper discussing future work.

2 VISION Infrastructure

VISION will develop an innovative infrastructure aiming at potentiating future WSNs with the capability of supporting intelligent audio, video and sensing services to be used by ubiquitous applications, with particular emphasis on real-time 3D video sensing.

The *VISION Infrastructure* is based on the layered architecture shown in Fig. 1 where both hardware (HW) and software (SW) components are depicted as gray and white boxes, respectively, ellipses represent different types of communication networks.

At the lowest level of Fig. 1 there is the WSN. In VISION, the WSN [2] is a specialized ad hoc network composed of different types of sensor nodes:

- *Reduced Functionality Devices* (RFDs): these are simplified nodes with low computational power, but conceived to be fully mobile, very low size and low power. They are able to manage simple environmental data acquisition and can communicate by low data rate links. RFDs collect information about the surrounding environment and they are interfaced with external sink nodes that issue queries about sensed data to the network as a whole.
- *Full Functionality Devices* (FFDs): nodes with a powerful computational power, low mobility and significant power consumption. They use a variable data rate according to the QoS. These devices support high-quality video sensing, and comunicate with a *gateway* (GW) through high data rate connections. They act as sink nodes for clusters of RFDs.
- *Gateways* (GWs) and *Servers*: nodes with a powerful computational power, the sink node may either store the data that can be processed off-line or forward them to a farther control unit, indicated as *Server*. Hence, the GW is equipped with radio interfaces to communicate with the underlying FFDs and RFDs nodes and by several communication interfaces based on the most common standards, like wi-fi, wi-max, GPRS, Ethernet.

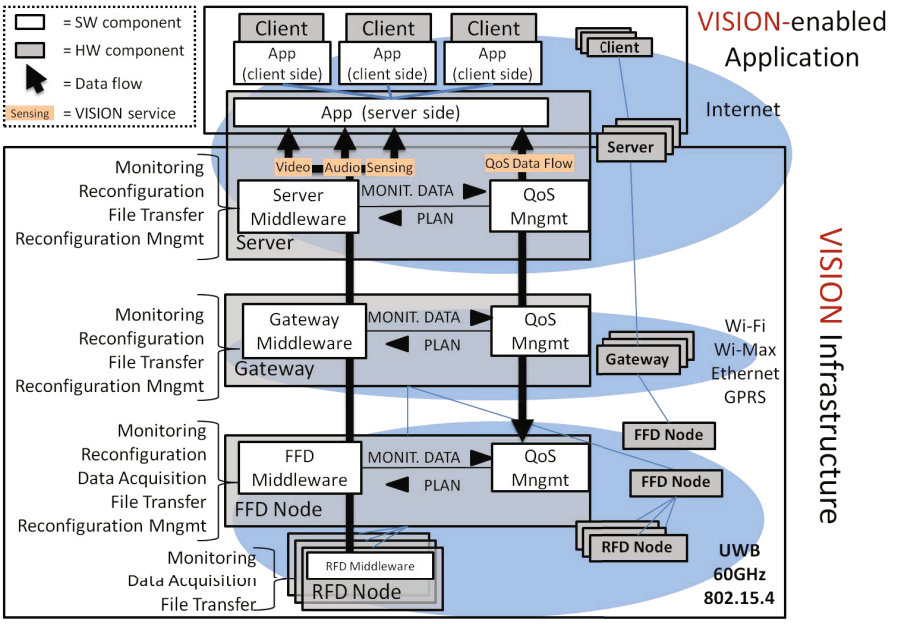


Fig. 1. VISION infrastructure

Software components of the VISION Infrastructure include middleware components that are specific of different hardware devices (RFDs, FFDs, GWs and Servers) and the components aimed at managing the Quality of Service (QoS).

The main functionalities of such software components (listed next to the middleware components in Fig. 1) can be grouped in

- **Device Specific Functionalities** that include: *monitoring* (i.e., the capability to access information regarding the device status); *reconfiguration* (i.e., the capability to execute device-specific reconfiguration plan); *data acquisition* from the external environment (e.g., temperature);
- **Data Control Functionalities** that comprehend: *reconfiguration management* (i.e., the capability to generate device-specific reconfiguration plan), and *file transfer* (i.e., providing video, audio, and sensing services to VISION-enabled applications);
- **QoS Control Functionalities** that generate a *reconfiguration plan* for the whole VISION infrastructure by accessing the monitored data.

3 Scenario: Communicative Impaired Children

This section presents a cognitive behavioural system where VISION technologies may apply. In particular, Section 3.1 describes the scenario, and presents the entities involved in the scenario as well as their interactions, Section 3.2 explains the context and the

assumptions under which the scenario works, and finally it summarizes the challenges we aim at addressing in the scenario as well as in the VISION infrastructure.

3.1 Description of the Scenario

People who have difficulties in communicating verbally, such as communicative impaired children (i.e. children with aphasia or dyslexia, general speech language impairments), send nonverbal messages that are not immediately captured by caregivers and parents, especially if they are not expert. For example, a child might appear calm and receptive to learning or enjoining in playing a game, while he/she is sad or angry or simply he/she dislikes the activity in which is involved. For this reason he/she is inattentive or may become aggressive for no reason. This scenario makes useless the attempt to teach or simply entertain the children.

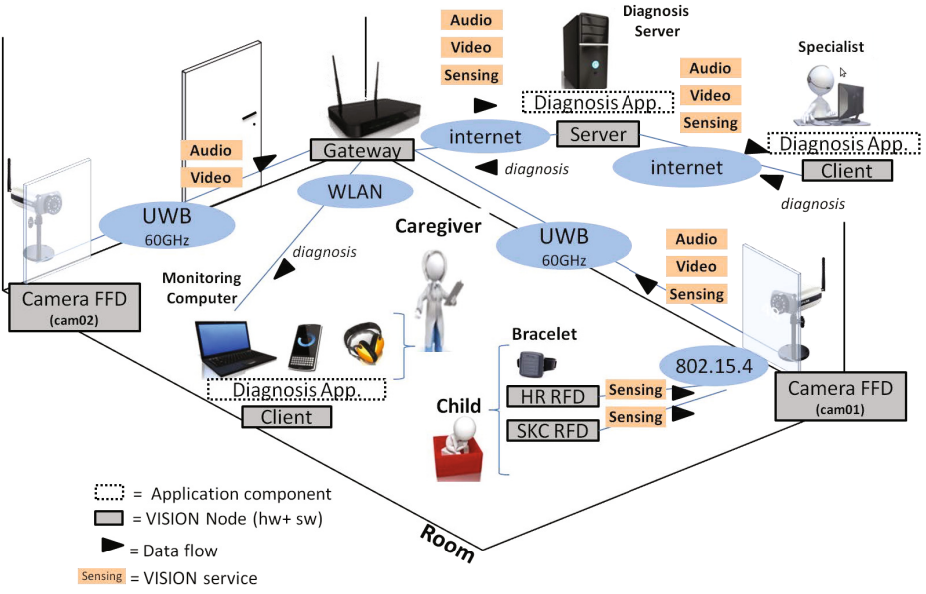


Fig. 2. Communicative impaired children scenario: overview

This fundamental communication problem can be addressed by enabling two main functionalities: (i) to collect longitudinal data of emotion-related expressive and physiological signals, in order to reveal the internal state that influence child's behaviour; (ii) to provide such information to caregivers to quickly intervene, in order to test the clinical and perceived effectiveness of the applied therapies in the treatments of childhood language disorders.

The functionality (i) is enabled if the room is equipped with hidden video sensors, the patient is equipped with physiological sensors analyzed in a body area network and hidden microphones are placed around. The functionality (ii) is enabled if the above

mentioned sensing units can transmit the captured data in real-time to a control unit, which forwards global status information, alerts and suggests suitable behavior and therapies to the caregiver. VISION is able to provide both functionalities in an optimized way because its wireless technology allows the best placement of sensors transparently to the impaired child, even on the his body; the broadband technology enables the real-time transfer to the control unit; and the middleware allows the best reliability of the system and the selection of the relevant data to be collected and transmitted.

The VISION equipment (see Fig. 2) is composed by: (i) one or more battery-powered *cameras* monitoring the behavioural activity of the *children*; (ii) software to process video sequences that should allow to capture and recognize emotional expressions by gestures, speech, and facial configurations (*Diagnosis App*). Additionally, some personalized physiological indicators, like heart-rate and skin-conductance sensors, i.e. hidden into wearable objects with software interfaces (e.g., a *bracelet*) may be used from the Diagnosis Application to define the child's daily arousal profile and to give feedback to *caregivers* and parents about how to interact with him.

Such a monitoring will facilitate the understanding of the children's physiological state and their behaviour, thus to allow reply to questions like "Which state helps the children best to maintain attention and focus for learning?". The goal is to assist the caregivers in improving the behavioural therapy, and making easier to track progresses and interventions for maximizing the learning process.

VISION can additionally provide artificial tools, such as virtual agents and avatars, that can help in maintaining the attentive focus of the children and enhance the quality of their learning engaging them in new activities as soon as the monitoring of their involvement in the current activity is understood by the system as degraded or not anymore attractive.

Participants and Interactions

In the proposed scenario a *caregiver* is teaching a dyslexic *child* to discriminate colors, letting him to play with colored cards and lists of words. The location is a *room* (i.e., at home, at a rehabilitation institute, or at hospital) that is equipped with two *cameras* in order to acquire recordings of the *child* facial expressions and body movements.

The *child* is wearing a *bracelet* (or any other non-invasive tool) which includes sensors for measuring heart rate (HR) and skin conductance (SKC). All the data collected by the cameras and the physiological indicator devices are sent to a multinetwork *gateway* that routes through the *Internet* towards a central *diagnosis server*. A VISION-enabled *diagnosis application* deployed on such a server is able to process the retrieved data both automatically or by requesting the help of a *remote psychologist*. Note that the location of this *diagnosis server* could be different from the room where the caregiver and the child are.

The processing should results in a series of data that describe and measure engagement level of the *child*, the quality of her/his learning, her/his emotional status. At the light of such outcomes, the diagnosis sent by the *diagnosis server*, through the *gateway* connected to *room*'s wireless local area network (*WLAN*), is routed to a local *monitoring computer*. Such a diagnosis:

Table 1. Participants of the Communicative impaired children scenario

Participant	Type	Localization	Role
Child	person	room	She/he is playing/learning
Caregiver	person	room	She/he is teaching the child
Specialist	person	outside room	She/he analyzes and combines sensed data if required by the diagnosis application
Camera (VISION FFD node)	resource	room	It acquires and sends video and audio streams
Gateway (VISION Gateway node)	resource	room	It routes sensed and application data from/to internal resources and from/to Internet
HR and SKC Sensor (VISION RFD node)	resource	on child	It acquires and sends physiological measures (heart rate and skin conductance)
Diagnosis Server (VISION Server)	resource	outside room	It processes the sensed data (audio, video, physiological measures)
Computer, headphone, smartphone (Client)	resource	room, on caregiver	It display diagnosis, sensed data, alarms and suggestions

- allows caregivers to visualize the real-time child's internal profiles, and raises alarms for events of interest in order to enable appropriate interventions (stop/change child's activity);
- provides some suggested reaction approaches on what is best to do to handle the current situation.

This latter output implicitly assumes that VISION is provided of an internal knowledge about the strengths that category of impaired children appear to have over typically developed ones. For example, it is known from the literature that dyslexic children have over non-dyslexic ones, extra creativity and stronger sensory receptors (tactile/touch) and therefore a multisensory structured teaching approach would maximize their learning processes.

Table 1 reports all the participants involved in the scenario listing the type of participant, their localization and the role they play.

3.2 Analysis of the Scenario

The context of the scenario is represented by a child that cannot verbally communicate in a proper way and she/he is involved in a learning task with a caregiver. The caregiver may not be an expert (like, e.g., a parent) and she/he cannot assess the child involvement and attentiveness for the given task, as well as cannot determine if the child is enjoining the playing/learning game, and more importantly she/he does not know what to do in case of alarming behaviours or loss of interest in the playing/learning game.

In the following list we schematically report the assumptions of the proposed scenario:

- The child and the caregiver are located in a room (at the child home, or a rehabilitation institute, or the hospital) in which cameras are able to follow and record the child movements;
- Cameras are appropriately mounted in a not visible way and they are connected through a wireless network;
- The caregiver is connected to the VISION system through her/his cell phone, and/or miniaturized wireless auriculares.
- The child is equipped with hidden, wireless and functional (not visible and/or not interpretable as physiological devices, such as bracelet-like) physiological sensors.
- The wireless local area network connects cameras and physiological sensors and transmits the collected signals to the VISION diagnosis server.
- The diagnosis server detects and analyzes the received information, exploits its knowledge unit in order to assess the results of the analysis, and sends feedback (in video and/or speech format) to the caregiver auriculares.
- The diagnosis server implements interventions on the monitor screen in the playing room, for example displaying avatars or a virtual characters that attract the child attention.

Table 2 schematically reports the challenges of the scenario under study (described in the first column in the table), and how such challenges can be addressed in the VISION infrastructure (described in the second column in the table).

In particular, the scenario requires a continuous video streaming provision that can result into several challenges to be addressed. In Table 2 we report two challenges related to cameras that if managed can guarantee the continuous video streaming provision in case the active camera fails or its battery reaches a critical level. VISION helps to manage both critical situation as we show in the next Section.

Indeed, in the first case (related to Fault tolerance for cameras) the VISION infrastructure provides redundant acquisition sources, and if a camera stops working it is possible to activate another camera, if available, thus to keep alive the video data acquisition service. This action is done automatically by the system, hence it is absolutely transparent to the child, who must not be aware of being recorded and monitored.

In the second case (related to the Resource awareness for Cameras) the VISION infrastructure monitors the battery level for the cameras. If it detects a low battery level for a camera then it can react by switching to a new one, if available, or by reducing the quality of data acquisition. Again, all this is done automatically and transparently to the users of the system.

4 VISION at Work on the Scenario

In this Section we detail how the VISION infrastructure works to provide some of its main functionalities through the use of UML Sequence Diagrams.

Fig. 3 shows normal workflow where middleware components running on VISION nodes (*cameras*, *gateway* and *server*) interact to provide the Video Streaming service. In

Table 2. Challenges for the scenario and their impact to the VISION infrastructure

Scenario challenge	VISION challenge
Fault tolerance for cameras.	The VISION infrastructure provides redundant acquisition sources; if a camera stops working it is possible to activate another camera, if available, thus to keep alive the video data acquisition.
Fault tolerance for body sensors.	The VISION infrastructure provides multiple body sensors; if a sensor stops working it is possible to activate another sensor, if available, thus to keep alive the heart rate and skin conductance data acquisition.
Fault-tolerance for the Wireless connection.	The VISION infrastructure provides reconfiguration mechanisms to manage resources. If VISION detects a degradation of the quality of the network connection then it can react by enabling a lower resolution for cameras or by reducing the number of active cameras.
Transparency for cameras (i.e., cameras are appropriately mounted in a not visible way).	The VISION infrastructure includes small sensor devices that can be hidden in unsuspicious objects thus do not interfere with the child perceptiveness of the context.
Transparency for Body sensors (i.e., body sensors are appropriately mounted in a not visible way).	The VISION infrastructure includes small body sensors that can be hidden in a bracelet thus do not interfere with the child perceptiveness of the context.
Resource awareness for Cameras.	If VISION detects a low battery level for a camera then it can react by switching to a new one, if available, or by reducing the quality of data acquisition.
Resource awareness for body sensors.	If VISION detects a low battery level for the body sensors then it can react by switching to new ones, if available, or by reducing the sampling frequency.

particular, in the communicative impaired children scenario (see Section 3) two hidden cameras (*cam01* and *cam02*) continuously send their video streams to a gateway (*gw*) that, at the same time, forwards such streams to the diagnosis server and monitors the status of VISION nodes. According to the challenges listed in Table 2, the VISION Infrastructure should exhibit resource awareness and fault tolerance capabilities.

The latter is illustrated in Fig. 4a where a failure occurs on *cam01*. Thanks to the continuos resource monitoring, the gateway is able to detect a failure resource signal. The latter triggers a reactive reconfiguration algorithm that results in a reconfiguration plan for the current system configuration. Such a plan, applied to the system configuration shown in Fig. 2, requires both starting a video stream from *cam02* (if not yet started) and re-establishing wireless connections using the 802.15.4 standard (see Fig. 2) between RFD sensors on the child's bracelet and *cam02*.

A resource-awareness scenario is shown in Fig. 4b. In this case *cam01* is able to monitor that charge level of its battery is under a certain treshold. A *low battery* signal is generated and received by the gateway thanks to resource monitoring. A proactive reconfiguration plan that optimize the energy consumption of *cam01* is generated and sent by the gateway. The affected camera, once received the plan reduce the resolution and/or frame rate of its video stream without interrupting the service.

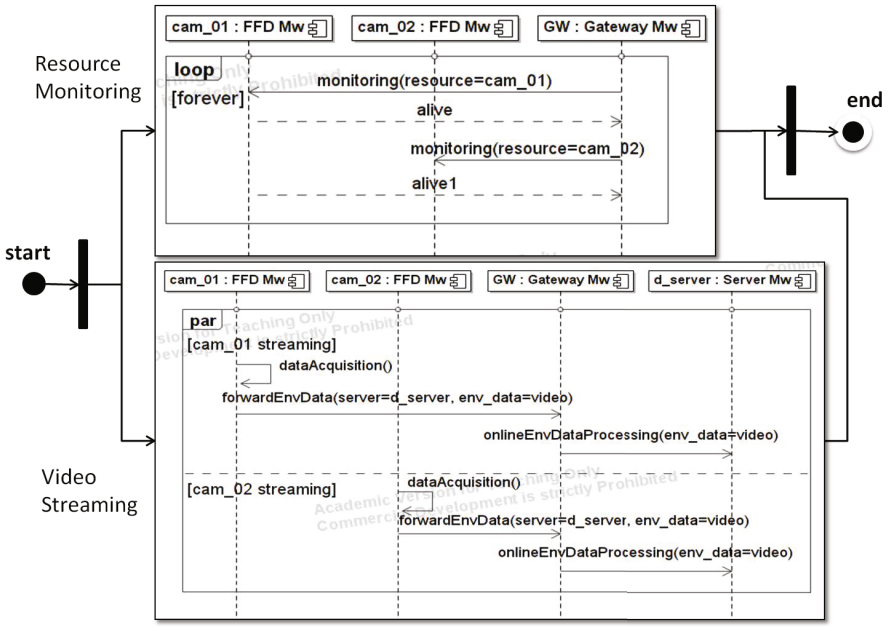


Fig. 3. VISION Monitoring and Video Streaming scenarios

5 Related Works

Recently, growing interest is focusing on cognitive systems [8] that are systems embedding and using psychological data, similar to the ways humans think and process information. They engage the functions of human recognition and increase one's cognitive capabilities helping, for example, people with difficulties on processing information.

A cognitive system to support and amuse elderly people in every day life is proposed in [3], where an infrastructure of wireless sensors is built in the kitchen and is networked with sensors placed over a mobile robot and over the involved elder person. The data collected by these WSNs are used by the robot to learn the fundamental gestures and movements humans make in a kitchen; and by the monitoring equipment to give the robot real-time feedback. In this case, the involved elder person wants to teach the companion robot how to behave, hence the infrastructure is visible and sensors are sometimes wire-connected. In our case, instead, we assume that the children are not aware that they are monitored, because this may influence the learning process. Since our scenario does not include a mobile robot, redundancy of points of observation is essential to achieve the required level of reliability of the system and to enable the re-configuration mechanisms that adapt the involved sensor asset to the needs of the instantaneous situation.

As proved in [3], cognitive systems can leverage WSN [2] to support the users to capture and elaborate context information, in a pervasive and ubiquitous way. The wireless technologies must be able to transparently sense the context in terms of audio, video

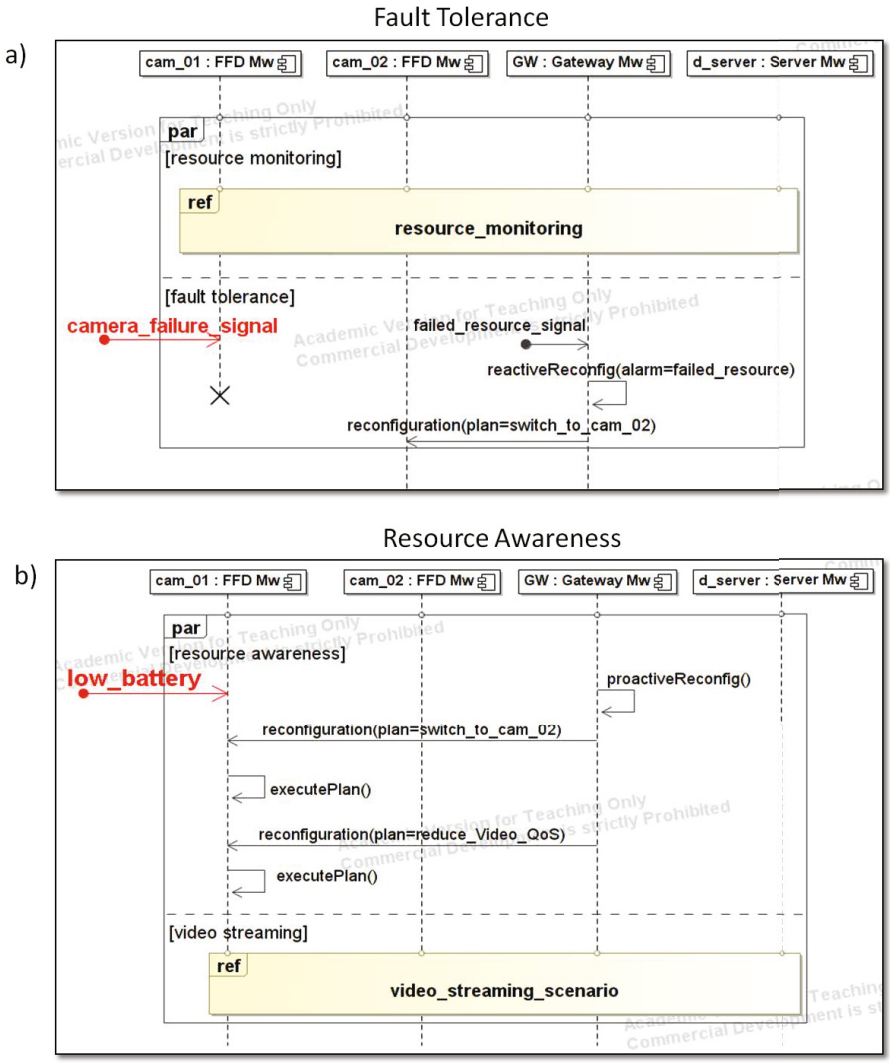


Fig. 4. a) Resource Awareness Scenario, b) Fault Tolerance Scenario

and other typical sensor data. Wireless Sensor Networks are suitable to host such information, since they are composed of a large number of sensor nodes that comprise one or more sensing units, a processor, a radio transceiver, and they are powered by an embedded battery. Sensors collect information about the surrounding environment (sensor field) and they are interfaced with external sink nodes that issue queries about sensed data.

As outlined in [5], the technical approach for developing cognitive systems involves these challenges: (i) the specification of the functionality and interfaces of the different cognitive systems; (ii) the completion of the development of the different cognitive sys-

tems including management algorithms and control channels; (iii) the realization of the necessary integration, making thus a step towards federation of numerous technologies, which will be a key for achieving the management of cognitive systems.

In particular, our paper works towards this latter goal, i.e. to meet the requirements for the effective management of a complex and heterogeneous infrastructure, based on WSN, through which users can obtain a variety of information at the best possible Quality of Service (QoS) levels, anytime, anywhere.

Currently, the envisaged applications are often discouraged by limitations in the sensor nodes power supply, communication bandwidth, processing capabilities and buffer size. As applications become more and more mission-critical (such as cognitive systems are), it is crucial that the collected sensor data are delivered to the sink within a specified time limit. Guaranteeing a certain quality of service (QoS) to a user or an application is difficult because of the unpredictable nature of the wireless link and the often unstable topology of the sensor network (due to node failure or mobility).

Very little research has been done in the field of QoS for WSNs [4,6] and many interesting research questions are still unanswered [7]. VISION will address most of the above mentioned limitations of current technology in order to design and prototype revolutionary HW/SW infrastructure able to provide intelligent sensing services, with particular emphasis on real-time 3D video, by properly combining the exploitation of resource/context-awareness of all system components with the use of a novel wireless technology, the 60 GHz ultra-wide band (UWB) radios, enabling broadband transmissions in WSNs.

6 Conclusions and Future Work

In this paper we have presented the VISION infrastructure and we have discussed how it can make easy the development of cognitive systems. In particular, we showed VISION at work on a system helping in the learning process of impaired children: the system is able to capture nonverbal messages sent by impaired children (i.e. children with aphasia or dyslexia, general speech language impairments), that are difficult to be immediately captured by caregivers and parents, especially if they are not expert. The system equipped with the VISION infrastructure can help the caregiver to quickly understand the children behavior, thus to react with suitable teaching activities.

From this experiment, we can conclude that the VISION infrastructure is a promising framework for cognitive behavioral systems that aim at recognizing a given status of the context, e.g., multi modal interactional information (such as facial expressions and body movements) in an automated way. The main innovations of VISION are: (i) relying in new emerging mm-waves and ultra-wideband wireless technologies; (ii) providing a reconfigurable middleware by applying the most advanced tools and techniques for context-aware systems, thus to enable the re-configurability of WSNs.

Currently the VISION system has been partially implemented, however in the future we aim at completing its development. The completion of the VISION system enables the support to cognitive behavioral systems, and different scenarios may benefit of the VISION enhanced technology. Our final purpose is to implement parts of the presented scenario as a demo for the project.

References

1. ERC Starting Independent Grant VISION, http://www.vision-ercproject.eu
2. Akyildiz, I.: Wireless sensor networks: a survey. Computer Networks 38(4), 393–422 (2002)
3. Beetz, M., Stulp, F., Radig, B., Bandouch, J., Blodow, N., Dolha, M., Fedrizzi, A., Jain, D., Klank, U., Kresse, I., Maldonado, A., Marton, Z., Mosenlechner, L., Ruiz, F., Rusu, R.B., Tenorth, M.: The assistive kitchen: A demonstration scenario for cognitive technical systems. In: The 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2008, pp. 1–8 (August 2008)
4. Chen, J., Díaz, M., Llopis, L., Rubio, B., Troya, J.M.: A survey on quality of service support in wireless sensor and actor networks: Requirements and challenges in the context of critical infrastructure protection. Journal of Network and Computer Applications 34(4), 1225–1239 (2011)
5. Dimitrakopoulos, G., Demestichas, P., Koenig, W.: Introduction of cognitive systems in the wireless world: Research achievements and future challenges for end-to-end efficiency. In: Future Network and Mobile Summit, pp. 1–9 (June 2010)
6. Cognitive Systems Research Journal 1-14, http://www.journals.elsevier.com/cognitive-systems-research/
7. Park, C., Rappaport, T.: Short-Range Wireless Communications for Next-Generation Networks: UWB, 60 GHz Millimeter-Wave WPAN, And ZigBee. IEEE Wireless Communications 14(4), 70–78 (2007)
8. Zhang, Q., Lee, M.: Emotion development system by interacting with human eeg and natural scene understanding. Cognitive Systems Research 14(1), 37–49 (2012); Cognitive Systems Research: Special Issue on Modeling and Application of Cognitive Systems

The Hourglass of Emotions

Erik Cambria[1], Andrew Livingstone[2], and Amir Hussain[3]

[1] Temasek Laboratories, National University of Singapore, Singapore
[2] Dept. of Psychology, University of Stirling, UK
[3] Dept. of Computing Science & Maths, University of Stirling, UK
cambria@nus.edu.sg, {a.g.livingstone,ahu}@stir.ac.uk
http://sentic.net

Abstract. Human emotions and their modelling are increasingly understood to be a crucial aspect in the development of intelligent systems. Over the past years, in fact, the adoption of psychological models of emotions has become a common trend among researchers and engineers working in the sphere of affective computing. Because of the elusive nature of emotions and the ambiguity of natural language, however, psychologists have developed many different affect models, which often are not suitable for the design of applications in fields such as affective HCI, social data mining, and sentiment analysis. To this end, we propose a novel biologically-inspired and psychologically-motivated emotion categorisation model that goes beyond mere categorical and dimensional approaches. Such model represents affective states both through labels and through four independent but concomitant affective dimensions, which can potentially describe the full range of emotional experiences that are rooted in any of us.

Keywords: Cognitive and Affective Modelling, NLP, Affective HCI.

1 Introduction

Emotions are an essential part of who we are and how we survive. They are complex states of feeling that result in physical and psychological reactions influencing both thought and behaviour. The study of emotions is one of the most confused (and still open) chapters in the history of psychology. This is mainly due to the ambiguity of natural language, which does not allow to describe mixed emotions in an unequivocal way. Love and other emotional words like anger and fear, in fact, are suitcase words (many different meanings packed in), not clearly defined and meaning different things to different people [1].

Hence, more than 90 definitions of emotions have been offered over the past century and there are almost as many theories of emotion, not to mention a complex array of overlapping words in our languages to describe them. Some categorisations include cognitive versus non-cognitive emotions, instinctual (from the amygdala) versus cognitive (from the prefrontal cortex) emotions, and also categorisations based on duration, as some emotions occur over a period of seconds (e.g., surprise), whereas others can last years (e.g., love).

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 144–157, 2012.

The James-Lange theory posits that emotional experience is largely due to the experience of bodily changes [2]. Its main contribution is the emphasis it places on the embodiment of emotions, especially the argument that changes in the bodily concomitants of emotions can alter their experienced intensity. Most contemporary neuroscientists endorse a modified James-Lange view, in which bodily feedback modulates the experience of emotion [3]. In this view, emotions are related to certain activities in brain areas that direct our attention, motivate our behaviour, and determine the significance of what is going on around us. Pioneering works by Broca [4], Papez [5], and MacLean [6] suggested that emotion is related to a group of structures in the centre of the brain called limbic system (or paleomammalian brain), which includes the hypothalamus, cingulate cortex, hippocampi, and other structures. More recent research, however, has shown that some of these limbic structures are not as directly related to emotion as others are, while some non-limbic structures have been found to be of greater emotional relevance [7].

Emotions are different Ways to Think [1] that our mind triggers to deal with different situations we face in our lives. Strong emotions can cause you to take actions you might not normally perform, or avoid situations that you generally enjoy. The affective aspect of cognition and communication, in fact, is recognised to be a crucial part of human intelligence and has been argued to be more fundamental in human behaviour and success in social life than intellect [8, 9]. Emotions influence cognition, and therefore intelligence, especially when this involves social decision-making and interaction. For this reason, human emotions and their modelling are increasingly understood to be a crucial aspect in the development of intelligent systems.

In particular, within sentic computing [10], a multi-disciplinary approach to opinion mining at the crossroads between affective computing and common sense computing, we developed a novel emotion categorisation model that allows to properly express the affective information associated with natural language text, for both emotion recognition and polarity detection tasks. A preliminary version of the model has already been used in some of our previous works [11–14], in which, however, no explicit motivations and details about the model were provided. The structure of the paper is as follows: Section 2 presents an overview of existing emotion categorisation models, Section 3 thoroughly explains motivations, peculiarities, and advantages of our model, Section 4, eventually, comprises concluding remarks and future directions.

2 Background

Philosophical studies on emotions date back to ancient Greeks and Romans. Following the early Stoics, for example, Cicero enumerated and organised the emotions into four basic categories: *metus* (fear), *aegritudo* (pain), *libido* (lust), and *laetitia* (pleasure). Studies on evolutionary theory of emotions, in turn, were initiated in the late 19th century by Darwin [15].

His thesis was that emotions evolved via natural selection and therefore have cross-culturally universal counterparts. In the early 1970s, Ekman found evidence

that humans share six basic emotions: happiness, sadness, fear, anger, disgust and surprise [16]. Few tentative efforts to detect non-basic affective states, such as fatigue, anxiety, satisfaction, confusion, or frustration, have been also made [17–22] (Table. 1). In 1980, Averill put forward the idea that emotions cannot be explained strictly on the basis of physiological or cognitive terms. Instead, he claimed that emotions are primarily social constructs; hence, a social level of analysis is necessary to truly understand the nature of emotion [23].

The relationship between emotion and language (and the fact that the language of emotion is considered a vital part of the experience of emotion) has been used by social constructivists and anthropologists to question the universality of Ekman's studies, arguably because the language labels he used to code emotions are somewhat US-centric. In addition, other cultures might have labels that cannot be literally translated to English (e.g., some languages do not have a word for fear [24]). For their deep connection with language and for the limitedness of the emotional labels used, all such categorical approaches usually fail to describe the complex range of emotions that can occur in daily communication.

The dimensional approach [25], in turn, represents emotions as coordinates in a multi-dimensional space. For both theoretical and practical reasons, more and more researchers like to define emotions according to two or more dimensions. An early example is Russell's circumplex model [26], which uses the dimensions of arousal and valence to plot 150 affective labels (Fig. 1). Similarly, Whissell considers emotions as a continuous 2D space whose dimensions are evaluation and activation [27]. The evaluation dimension measures how a human feels, from positive to negative. The activation dimension measures whether humans are more or less likely to take some action under the emotional state, from active to passive (Fig. 2).

In her study, Whissell assigns a pair of values <activation, evaluation> to each of the approximately 9,000 words with affective connotations that make up her Dictionary of Affect in Language. Another bi-dimensional model is Plutchik's wheel of emotions, which offers an integrative theory based on evolutionary principles [28]. Following Darwin's thought, the functionalist approach to emotions

Table 1. Some existing definition of basic emotions. The most widely adopted model for affect recognition is Ekman's, although is one of the poorest in terms of number of emotions.

Author	#Emotions	Basic Emotions
Ekman	6	anger, disgust, fear, joy, sadness, surprise
Parrot	6	anger, fear, joy, love, sadness, surprise
Frijda	6	desire, happiness, interest, surprise, wonder, sorrow
Plutchik	8	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins	9	desire, happiness, interest, surprise, wonder, sorrow
Matsumoto	22	joy, anticipation, anger, disgust, sadness, surprise, fear, acceptance, shy, pride, appreciate, calmness, admire, contempt, love, happiness, exciting, regret, ease, discomfort, respect, like

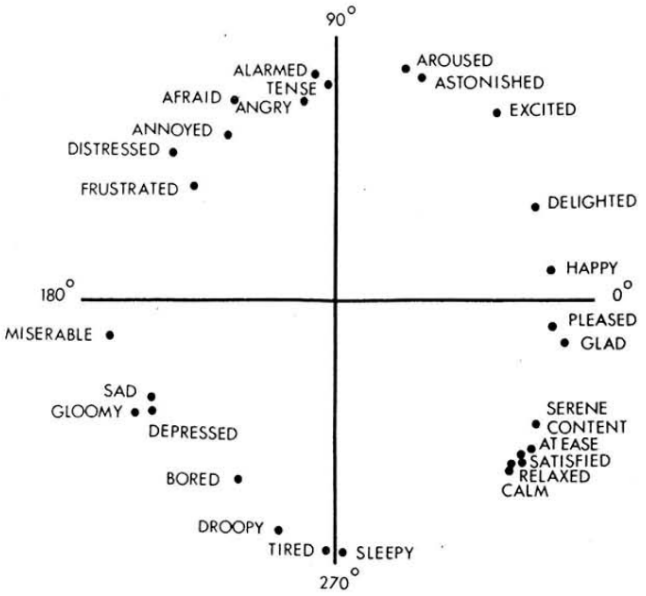


Fig. 1. Russell's circumplex model is one of the earlies examples of dimensional emotion representations. In the snippet, direct circular scaling coordinates are provided for 28 affect words.

holds that emotions have evolved for a particular function, such as to keep the subject safe [29, 30]. Emotions are adaptive as they have a complexity born of a long evolutionary history and, although we conceive emotions as feeling states, Plutchik says the feeling state is part of a process involving both cognition and behaviour and containing several feedback loops. In 1980, he created a wheel of emotions that consisted of 8 basic emotions and 8 advanced emotions each composed of 2 basic ones. In such model, the vertical dimension represents intensity and the radial dimension represents degrees of similarity among the emotions.

Besides bi-dimensional approaches, a commonly used set for emotion dimension is the <arousal, valence, dominance> set, which is known in the literature also by different names, including <evaluation, activation, power> and <pleasure, arousal, dominance> [31]. Recent evidence suggests there should be a fourth dimension: Fontaine et al. report consistent results from various cultures where a set of four dimensions is found in user studies, namely <valence, potency, arousal, unpredictability> [32].

Dimensional representations of affect are attractive mainly because they provide a way of describing emotional states that is more tractable than using words. This is of particular importance when dealing with naturalistic data, where a wide range of emotional states occurs. Similarly, they are much more able to deal with non-discrete emotions and variations in emotional states over time [33], since in such cases changing from one universal emotion label to another would not make much sense in real life scenarios.

Dimensional approaches, however, have a few limitations. Although the dimensional space allows to compare affect words according to their reciprocal distance, it usually does not allow to make operations between these, e.g., for studying compound emotions. Most dimensional representations, moreover, do not model the fact that two or more emotions may be experienced at the same time. Eventually, all such approaches work at word level, which makes them unable to grasp the affective valence of multiple-word concepts.

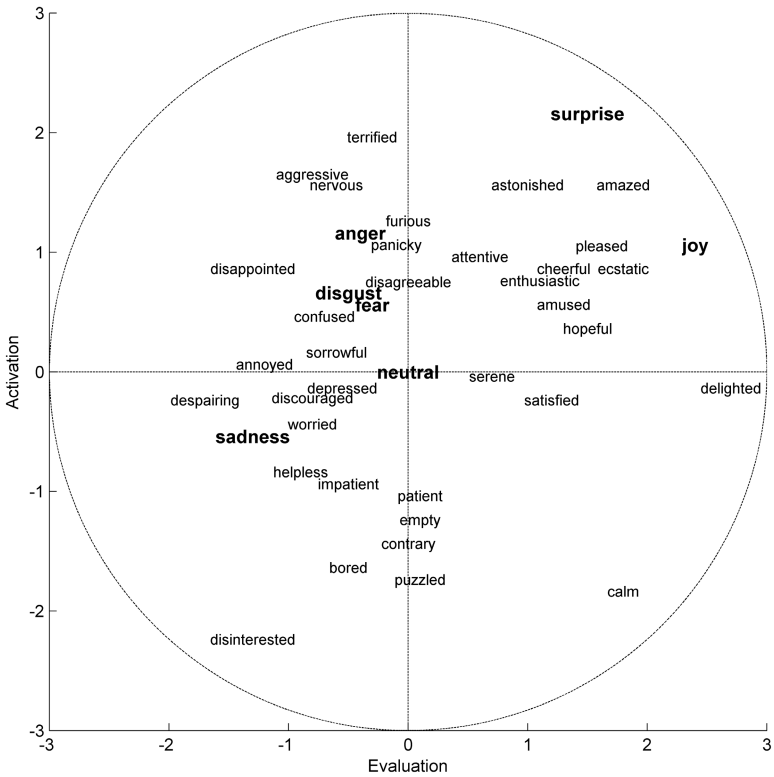


Fig. 2. Whissell's model is a bi-dimensional representation of emotions, in which words from the Dictionary of Affect in Language are displayed. The diagram shows the position of some of these words in the <activation, evaluation> space.

3 The Hourglass Model

The Hourglass of Emotions is an affective categorisation model primarily inspired by Plutchik's studies on human emotions [28]. It reinterprets Plutchik's model by organising primary emotions around four independent but concomitant dimensions, whose different levels of activation make up the total emotional state

of the mind. The main motivation for the design of the model is the concept-level inference of the cognitive and affective information associated with text. Such faceted information is needed, within sentic computing, for a feature-based sentiment analysis, where the affective common sense knowledge associated with natural language opinions has to be objectively assessed.

Therefore, the Hourglass model systematically excludes what are variously known as self-conscious or moral emotions such as pride, guilt, shame, embarrassment, moral outrage, or humiliation [34–37]. Such emotions, in fact, may still present something of a blind spot for models rooted in basic emotions, because they are by definition contingent on subjective moral standards. The distinction between guilt and shame, for example, is based in the attribution of negativity to the self or to the act. So, guilt arises when believing to have done a bad thing, and shame arises when thinking to be a bad person. This matters because in turn, these emotions have been shown to have different consequences in terms of action tendencies. Likewise, an emotion such as *schadenfreude* is essentially a form of pleasure, but it is crucially different from pride or happiness because of the object of the emotion (the misfortune of another that is not caused by the self), and the resulting action tendency (do not express).

However, since the Hourglass model currently focuses on the objective inference of affective information associated with natural language opinions, appraisal-based emotions are not taken into account within the present version of the model. Several affect recognition and sentiment analysis systems [38–44] are based on different emotion categorisation models, which generally comprise a relatively small set of categories (Table 2). The Hourglass of Emotions, in turn, allows classifying affective information both in a categorical way (according to a wider number of emotion categories) and in a dimensional format (which facilitates comparison and aggregation).

3.1 A Novel Cognitive Model for the Representation of Affect

The Hourglass of Emotions is a brain-inspired and psychologically-motivated model based on the idea that the mind is made of different independent resources and that emotional states result from turning some set of these resources on and turning another set of them off [1]. Each such selection changes how we think by changing our brain's activities: the state of anger, for example, appears to select a set of resources that help us react with more speed and strength while also suppressing some other resources that usually make us act prudently. Evidence of this theory is also given by several fMRI experiments showing that there is a distinct pattern of brain activity that occurs when people are experiencing different emotions.

Zeki and Romaya, for example, investigated the neural correlates of hate with an fMRI procedure [46]. In their experiment, people had their brains scanned while viewing pictures of people they hated. The results showed increased activity in the medial frontal gyrus, right putamen, bilaterally in the premotor cortex, in the frontal pole, and bilaterally in the medial insula of the human brain. Also the activity of emotionally enhanced memory retention can be linked to

Table 2. An overview of recent model-based affect recognition and sentiment analysis systems. Studies are divided by techniques applied, number of categories of the emotion categorisation model adopted, corpora and knowledge base used.

Study	Techniques	#Categories	Corpora	Knowledge Base
[40]	NB, SVM	2	Political articles	None
[41]	LSA, MLP, NB, KNN	3	Dialogue turns	ITS interaction
[44]	Cohesion indices	4	Dialogue logs	ITS interaction
[42]	VSM, NB, SVM	5	ISEAR	ConceptNet
[43]	WN presence, LSA	6	News stories	WNA
[38]	WN presence	6	Chat logs	WNA
[39]	Winnow linear, C4.5	7	Children stories	None
[45]	VSM, KNN	24	LiveJournal	ConceptNet, WNA
[11]	VSM, k-means	24	YouTube, LiveJournal	ConceptNet, WNA, HEO
[14]	VSM, k-means	24	LiveJournal, PatientOpinion	ConceptNet, WNA
[12]	VSM, k-medoids	24	Twitter, LiveJournal, PatientOpinion	ConceptNet, Probase

human evolution [47]. During early development, in fact, responsive behaviour to environmental events is likely to have progressed as a process of trial and error. Survival depended on behavioural patterns that were repeated or reinforced through life and death situations. Through evolution, this process of learning became genetically embedded in humans and all animal species in what is known as 'fight or flight' instinct [48].

The primary quantity we can measure about an emotion we feel is its strength. But, when we feel a strong emotion, it is because we feel a very specific emotion. And, conversely, we cannot feel a specific emotion like fear or amazement without that emotion being reasonably strong. For such reasons, the transition between different emotional states is modelled, within the same affective dimension, using the function $G(x) = -\frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$, for its symmetric inverted bell curve shape that quickly rises up towards the unit value (Fig. 3). In particular, the function models how the level of activation of each affective dimension varies from the state of 'emotional void' (null value) to the state of 'heightened emotionality' (unit value). Justification for assuming that the Gaussian function (rather than a step or simple linear function) is appropriate for modelling the variation of emotion intensity is based on research into the neural and behavioural correlates of emotion, which are assumed to indicate emotional intensity in some sense.

In fact, nobody genuinely knows what function subjective emotion intensity follows, because it has never been truly or directly measured [49]. For example, the so-called Duchenne smile (a genuine smile indicating pleasure) is characterised by smooth onset, increasing to an apex, and a smooth, relatively lengthy offset [50]. More generally, Klaus Scherer has argued that emotion is a process characterised by non-linear relations among its component elements - especially

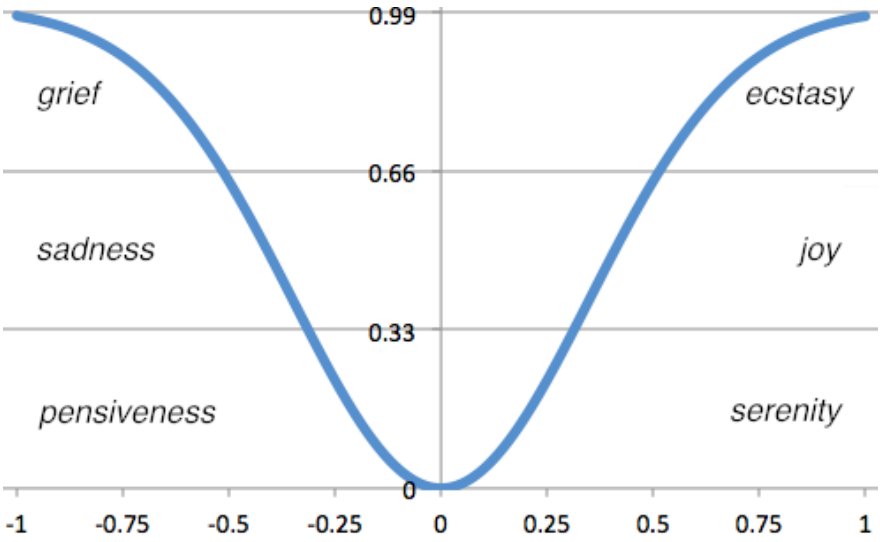


Fig. 3. The Pleasantness emotional flow. Within each affective dimension, the passage from a sentic level to another is regulated by a Gaussian function that models how stronger emotions induce higher emotional sensitivity.

physiological measures, which typically look Gaussian [51]. Emotions, in fact, are not linear [28]: the stronger the emotion, the easier it is to be aware of it.

Mapping the space of positive and negative primary emotions according to $G(x)$ leads to a hourglass shape (Fig. 4). It is worth to note that, in the model, the state of 'emotional void' is a-dimensional, which contributes to determine the hourglass shape. Total absence of emotion, in fact, can be associated with the total absence of reasoning (or, at least, consciousness) [52], which is not an envisaged mental state as, in human mind, there is never nothing going on.

3.2 A Model for Affective HCI

The Hourglass of Emotions can be exploited in the context of HCI to measure how much respectively: the user is amused by interaction modalities (Pleasantness), the user is interested in interaction contents (Attention), the user is comfortable with interaction dynamics (Sensitivity), the user is confident in interaction benefits (Aptitude).

Each affective dimension, in particular, is characterised by six levels of activation (measuring the strength of an emotion), termed 'sentic levels', which represent the intensity thresholds of the expressed/perceived emotion. These levels are also labelled as a set of 24 basic emotions [28], six for each of the affective dimensions, in a way that allows the model to specify the affective information associated with text both in a dimensional and in a discrete form (as shown in Table 3). The dimensional form, in particular, is called 'sentic vector'

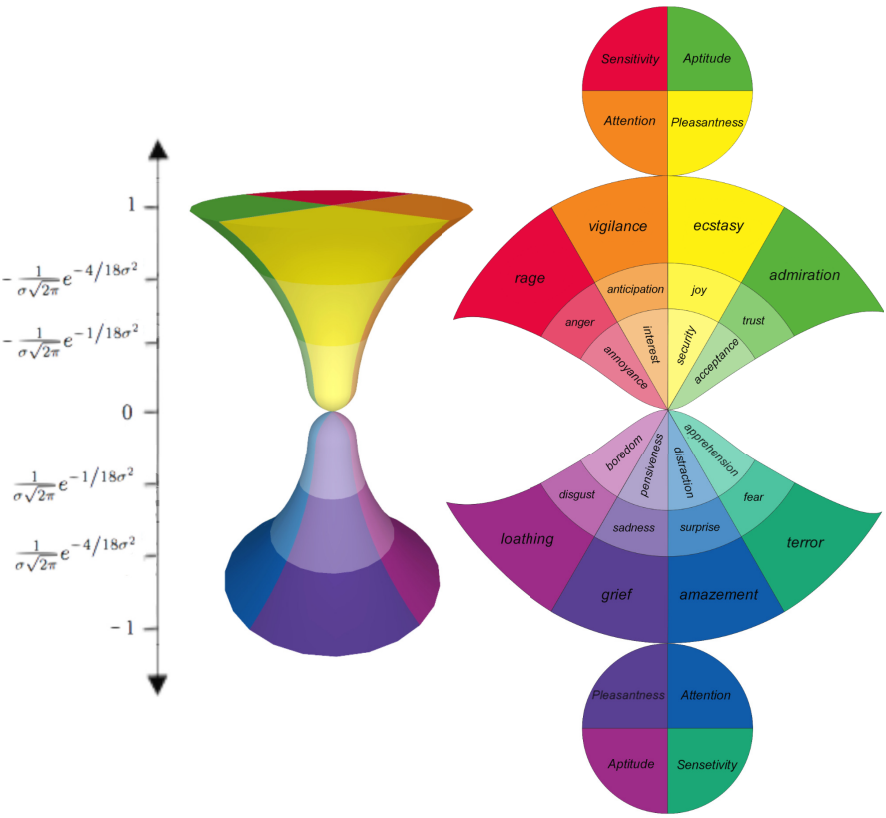


Fig. 4. The 3D model and the net of the Hourglass of Emotions: since affective states are represented according to their strength (from strongly positive to null to strongly negative), the model assumes a hourglass shape

and it is a four-dimensional *float* vector that can potentially synthesize the full range of emotional experiences in terms of Pleasantness, Attention, Sensitivity and Aptitude. In the model, the vertical dimension represents the intensity of the different affective dimensions, while the radial dimension models the activation of different emotional configurations, resembling Minsky's *k*-lines [53].

The model follows the pattern used in colour theory and research in order to obtain judgements about combinations, i.e., the emotions that result when two or more fundamental emotions are combined, in the same way that red and blue make purple. Hence, some particular sets of sentic vectors have special names as they specify well-known compound emotions (Fig. 5).

For example, the set of sentic vectors with a level of Pleasantness $\in$ [G(2/3), G(1/3)), i.e., joy, a level of Aptitude $\in$ [G(2/3), G(1/3)), i.e., trust, and a minor magnitude of Attention and Sensitivity, are called 'love sentic vectors' since they specify the compound emotion of love (Table 4). More complex emotions can be

Table 3. The sentic levels of the Hourglass model: each affective dimension contains six different levels of activation characterised by both a categorical and dimensional form

Interval	Pleasantness	Attention	Sensitivity	Aptitude
[G(1), G(2/3))	ecstasy	vigilance	rage	admiration
[G(2/3), G(1/3))	joy	anticipation	anger	trust
[G(1/3), G(0))	serenity	interest	annoyance	acceptance
(G(0), –G(1/3)]	pensiveness	distraction	apprehension	boredom
(–G(1/3), –G(2/3)]	sadness	surprise	fear	disgust
(–G(2/3), –G(1)]	grief	amazement	terror	loathing

synthesised by using three, or even four, sentic levels, e.g., joy + trust + anger = jealousy. Therefore, analogous to the way primary colours combine to generate different colour gradations (and even colours we do not have a name for), the primary emotions of the Hourglass model can blend to form the full range of emotional experiences that are rooted in anyone. Beyond emotion detection, the Hourglass model is also used for polarity detection tasks. Since polarity is strongly connected to attitudes and feelings, in fact, it can be defined in term of the four affective dimensions:

$$p = \sum_{i=1}^{N} \frac{Pleasantness(c_i) + |Attention(c_i)| - |Sensitivity(c_i)| + Aptitude(c_i)}{3N}$$

where c_i is an input concept, N the total number of concepts, and 3 the normalisation factor (as the Hourglass dimensions are defined as float $\in$ [-1,+1]).

Table 4. The second-level emotions generated by pairwise combination of the sentic levels of the Hourglass model. Different concomitant levels of activation give birth to different kinds of compound emotions, e.g., love, frustration, and anxiety.

	Attention>0	Attention<0	Aptitude>0	Aptitude<0
Pleasantness>0	optimism	frivolity	love	gloat
Pleasantness<0	frustration	disapproval	envy	remorse
Sensitivity>0	aggressiveness	rejection	rivalry	contempt
Sensitivity<0	anxiety	awe	submission	coercion

In the formula, Attention is taken in absolute value since both its positive and negative intensity values correspond to positive polarity values (e.g., surprise is negative in the sense of lack of Attention but positive from a polarity point of view). Similarly, Sensitivity is taken in negative absolute value since both its positive and negative intensity values correspond to negative polarity values (e.g., anger is positive in the sense of level of activation of Sensitivity but negative in terms of polarity).

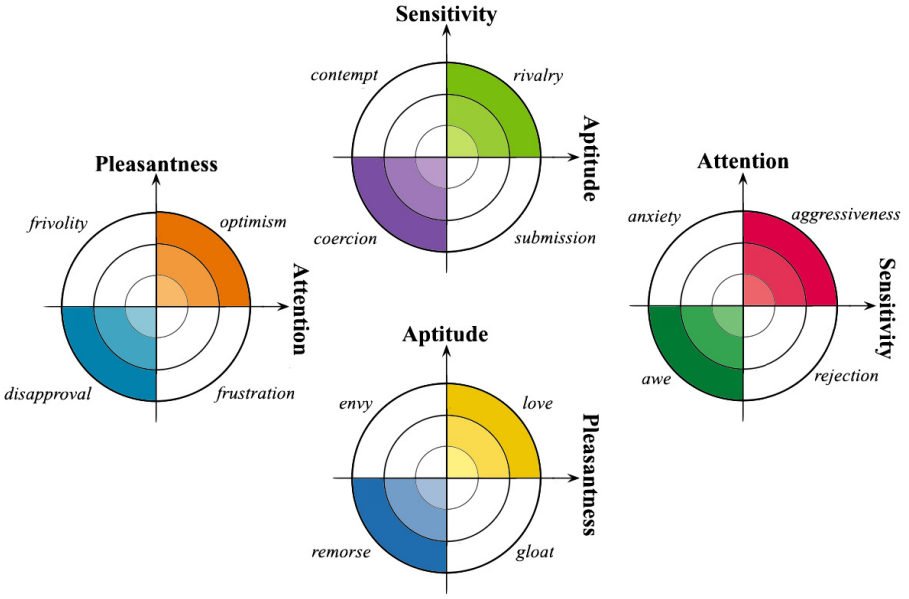


Fig. 5. Hourglass compound emotions of second level: by combining basic emotions pairwise it is possible to obtain complex emotions resulting from the activation of two of the four affective dimensions

4 Conclusion

Affective neuroscience and twin disciplines have clearly demonstrated how emotions and intelligence are strictly connected. Therefore, in order to enhance intelligent system processing and reasoning, it is necessary to provide machines with emotional models for time-critical decision enforcement.

Moreover, technology is increasingly used to observe human-to-human interactions, e.g., customer frustration monitoring in call centre applications. In such contexts, it is necessary to provide a suitable representation of emotional information, which should make the concepts and descriptions developed in the affective sciences available for use in technological contexts.

In this work, we developed the Hourglass of Emotions, a novel biologically-inspired and psychologically-motivated emotion categorisation model that goes beyond mere categorical and dimensional approaches. Such model represents affective states both through labels and through four independent but concomitant affective dimensions, which can potentially describe the full range of emotional experiences that are rooted in any of us.

In the future, we will be exploiting the model for the development of emotion-sensitive systems in different fields, in order to explore how much the model is generalisable and suitable for potentially any affective computing application.

We also plan to further modify the model in order to better represent compound emotions and to include the description of appraisal-based emotions.

Acknowledgements. We would like to thank Joseph Lyons, Integrator at Sitekit Solutions Ltd., for the precious help in the design and refinement of the Hourglass model.

References

1. Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon & Schuster (2006)
2. James, W.: What is an emotion? Mind 34, 188–205 (1884)
3. Dalgleish, T.: The emotional brain. Nature: Perspectives 5, 582–589 (2004)
4. Broca, P.: Anatomie comparée des circonvolutions cérébrales: Le grand lobe limbique. Rev. Anthropol. 1, 385–498 (1878)
5. Papez, J.: A proposed mechanism of emotion. Neuropsychiatry Clin. Neurosci. 7, 103–112 (1937)
6. Maclean, P.: Psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). Electroencephalogr Clin. Neurophysiol. suppl.4, 407–418 (1952)
7. Ledoux, J.: Synaptic Self. Penguin Books (2003)
8. Vesterinen, E.: Affective computing. In: Digital Media Research Seminar, Helsinki (2001)
9. Pantic, M.: Affective computing. In: Encyclopedia of Multimedia Technology and Networking, vol. 1, pp. 8–14. Idea Group Reference (2005)
10. Cambria, E., Hussain, A.: Sentic Computing: Techniques, Tools, and Applications. Springer, Dordrecht (2012)
11. Cambria, E., Grassi, M., Hussain, A., Havasi, C.: Sentic computing for social media marketing. Multimedia Tools and Applications 59(2), 557–577 (2012), http://dx.doi.org/10.1007/s11042-011-0815-0
12. Cambria, E., Song, Y., Wang, H., Hussain, A.: Isanette: A common and common sense knowledge base for opinion mining. In: ICDM, Vancouver, pp. 315–322 (2011)
13. Cambria, E., Havasi, C., Hussain, A.: SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In: FLAIRS, Marco Island, pp. 202–207 (2012)
14. Cambria, E., Benson, T., Eckl, C., Hussain, A.: Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. Expert Systems with Applications 39(12), 10533–10543 (2012), http://dx.doi.org/10.1016/j.eswa.2012.02.120
15. Charles, D.: The Expression of the Emotions in Man and Animals. John Murray (1872)
16. Ekman, P., Dalgleish, T., Power, M.: Handbook of Cognition and Emotion. Wiley, Chichester (1999)
17. Scherer, K.: Psychological models of emotion. The Neuropsychology of Emotion, 137–162 (2000)
18. Parrott, W.: Emotions in Social Psychology. Psychology Press (2001)
19. Prinz, J.: Gut Reactions: A Perceptual Theory of Emotion. Oxford University Press (2004)

20. Douglas-Cowie, E.: Humaine deliverable d5g: Mid term report on database exemplar progress. Technical report, Information Society Technologies (2006)
21. Kapoor, A., Burleson, W., Picard, R.: Automatic prediction of frustration. International Journal of Human-Computer Studies 65, 724–736 (2007)
22. Castellano, G., Kessous, L., Caridakis, G.: Multimodal emotion recognition from expressive faces, body gestures and speech. In: Doctoral Consortium of ACII, Lisbon (2007)
23. Averill, J.: A constructivist view of emotion. Emotion: Theory, Research and Experience, pp. 305–339 (1980)
24. Russell, J.: Core affect and the psychological construction of emotion. Psychological Rev. 110, 145–172 (2003)
25. Osgood, C., Suci, G., Tannenbaum, P.: The Measurement of Meaning. University of Illinois Press (1957)
26. Russell, J.: Affective space is bipolar. Journal of Personality and Social Psychology 37, 345–356 (1979)
27. Whissell, C.: The dictionary of affect in language. Emotion: Theory, Research, and Experience 4, 113–131 (1989)
28. Plutchik, R.: The nature of emotions. American Scientist 89(4), 344–350 (2001)
29. Frijda, N.: The laws of emotions. American Psychologist 43(5) (1988)
30. Freitas, A., Castro, E.: Facial expression: The effect of the smile in the treatment of depression. empirical study with portuguese subjects. In: Emotional Expression: The Brain and The Face, pp. 127–140. University Fernando Pessoa Press (2009)
31. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology 14(4), 261–292 (1996)
32. Fontaine, J., Scherer, K., Roesch, E., Ellsworth, P.: The world of emotions is not two-dimensional. Psychological Science 18(12), 1050–1057 (2007)
33. Cochrane, T.: Eight dimensions for the emotions. Social Science Information 48(3), 379–420 (2009)
34. Lazarus, R.: Emotion and Adaptation. Oxford University Press, New York (1991)
35. Lewis, M.: Self-conscious emotions: Embarrassment, pride, shame, and guilt. In: Handbook of Cognition and Emotion, vol. 2, pp. 623–636. Guilford Press (2000)
36. Scherer, K., Shorr, A., Johnstone, T.: Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, Canary (2001)
37. Tracy, J., Robins, R., Tangney, J.: The Self-Conscious Emotions: Theory and Research. The Guilford Press (2007)
38. Ma, C., Osherenko, A., Prendinger, H., Ishizuka, M.: A chat system based on emotion estimation from text and embodied conversational messengers. In: Int'l Conf. Active Media Technology, pp. 546–548 (2005)
39. Alm, C., Roth, D., Sproat, R.: Emotions from text: Machine learning for text-based emotion prediction. In: HLT/EMNLP, pp. 347–354 (2005)
40. Lin, W., Wilson, T., Wiebe, J., Hauptmann, A.: Which side are you on? identifying perspectives at the document and sentence levels. In: Conference on Natural Language Learning, pp. 109–116 (2006)
41. D'Mello, S., Craig, S., Sullins, J., Graesser, A.: Predicting affective states expressed through an emote-aloud procedure from autotutor's mixed-initiative dialogue. Int'l J. Artificial Intelligence in Education 16, 3–28 (2006)
42. Danisman, T., Alpkocak, A.: Feeler: Emotion classification of text using vector space model. In: AISB (2008)
43. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: ACM Symp. Applied Computing, pp. 1556–1560 (2008)

44. D'Mello, S., Dowell, N., Graesser, A.: Cohesion relationships in tutorial dialogue as predictors of affective states. In: Proceeedings of Conf. Artificial Intelligence in Education, pp. 9–16 (2009)
45. Grassi, M., Cambria, E., Hussain, A., Piazza, F.: Sentic web: A new paradigm for managing social media affective information. Cognitive Computation 3(3), 480–489 (2011)
46. Zeki, S., Romaya, J.: Neural correlates of hate. PloS One 3(10), 35–56 (2008)
47. Cahill, L., McGaugh, J.: A novel demonstration of enhanced memory associated with emotional arousal. Consciousness and Cognition 4(4), 410–421 (1995)
48. Bradford Cannon, W.: Bodily Changes in Pain, Hunger, Fear and Rage: An Account of Recent Researches into the Function of Emotional Excitement. Appleton Century Crofts (1915)
49. Barrett, L.: Solving the emotion paradox: Categorization and the experience of emotion. Personality and Social Psychology Review 10(1), 20–46 (2006)
50. Krumhuber, E., Kappas, A.: Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. Journal of Nonverbal Behavior 29(1), 3–24 (2005)
51. Lewis, M., Granic, I.: Emotion, Development, and Self-Organization: Dynamic Systems Approaches to Emotional Development. Cambridge University Press (2002)
52. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper Perennial (1991)
53. Minsky, M.: The Society of Mind. Simon and Schuster, New York (1986)

A Naturalistic Database of Thermal Emotional Facial Expressions and Effects of Induced Emotions on Memory

Anna Esposito[1], Vincenzo Capuano[1], Jiri Mekyska[2], and Marcos Faundez-Zanuy[3]

[1] Second University of Naples, Department of Psychology, Caserta, and IIASS, Italy
[2] Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Telecommunications, Brno, Czech Republic
[3] EUP Mataró, Avda. Puig i Cadafalch 101, 08303 Mataró (Barcelona), Spain
iiass.annaesp@tin.it, vincenzo.capuano85@gmail.com
xmekys01@stud.feec.vutbr.cz, faundez@eupmt.es

Abstract. This work defines a procedure for collecting naturally induced emotional facial expressions through the vision of movie excerpts with high emotional contents and reports experimental data ascertaining the effects of emotions on memory word recognition tasks. The induced emotional states include the four basic emotions of sadness, disgust, happiness, and surprise, as well as the neutral emotional state. The resulting database contains both thermal and visible emotional facial expressions, portrayed by forty Italian subjects and simultaneously acquired by appropriately synchronizing a thermal and a standard visible camera. Each subject's recording session lasted 45 minutes, allowing for each mode (thermal or visible) to collect a minimum of 2000 facial expressions from which a minimum of 400 were selected as highly expressive of each emotion category. The database is available to the scientific community and can be obtained contacting one of the authors. For this pilot study, it was found that emotions and/or emotion categories do not affect individual performance on memory word recognition tasks and temperature changes in the face or in some regions of it do not discriminate among emotional states.

Keywords: Database, Emotions, Thermal image, Naturalistic, Memory word recognition tasks.

1 Introduction

The testing of competitive algorithms through data shared by dozens of research laboratories is a milestone for getting significant technological advances [9]. Shared databases allow to validate and develop new algorithms, as well as assess their performance in order to select the most excellent for a given application. The advancement of the pattern recognition research community is measured on the performance obtained by the proposed pattern recognition systems on benchmark databases in fields such as biometrics, optical character recognition, medical images, object recognition, etc.

A challenging research topic in the field of Human-Machine Interaction is the analysis and recognition of emotional facial expressions. This is because recognizing

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 158–173, 2012.

faces (and in particular emotional faces) under gross environmental variations (such as the quality of the camera, light variations etcetera) and in real time remains a problem largely unsolved [3].

The collection and distribution of databases is a time-resource-consuming task, requiring experience and care both in the content design and the acquisition protocol. After the data collection, additional efforts are typically dedicated to supervise, annotate, label, error correct and document the collected data. In addition, a set of legal requirements have to be addressed, including consent forms to be signed by the donators and operational security measures as instructed by the data protection authorities. Finally, the distribution of the database involves intellectual property rights and maintenance issues.

When it comes to emotional facial expressions, according to the classical literature (largely debated but not yet superseded [7]) there are only 6 emotional categories[1] labelled as happiness, sadness, anger, fear, surprise, and disgust. However, the limited number of classes does not simplify the collection of a database of emotional facial expressions, due to the intrinsic difficulty to dispose of natural and spontaneous emotional samples from a significant amount of people. In order to collect such data there are mainly three procedures:

a) Recordings of spontaneous manifestations of emotional feelings. Generally this can be done by collecting video-recordings of subjects in their everyday activity, such as shopping, meeting, etcetera. The main drawback in such scenarios is the lack of control and therefore, a high amount of variability in the data, as well as the presence of few strikingly clear instances of episodic emotions. Cowie et al. [4], after the analysis of the Belfast naturalistic database, containing highly emotional talk-show recordings, showed that clear-cut emotional episodes were unexpectedly rare in such scenarios.
b) Recordings of subjects asked to simulate a specific facial emotional expression. Generally they are professional actors. However, although skilled actors can be convincing, it could be argued that they are not really experiencing the portrayed emotion, but a stylized version of the natural one, and therefore, a different set of facial features may be needed for their description. Batliner et al. [2] demonstrated that vocal signs of emotionality used by an actor simulating a particular human-machine interaction were different from, and much simpler than, those produced by people genuinely engaged in it.
c) Recordings of induced emotional states: to this aim there exists various emotion induction techniques. Some include the listening to emotional musical expressions, the watching of pictures and movies with highly emotional contents, as well as the playing of specially designed games. The advantage achieved in such scenarios is a higher situational control and thus, a major reliability of the collected data and the associated measurements.

[1] Not all the authors agree on these 6 (see [11] as an example).

An excellent overview of the different existing databases for the automatic modelling of emotional states is reported in [4].

According to the acquisition procedure, emotional databases can be split into four modalities: audio (typical measurements over speech signals are prosody, voice quality, timing, etc.), photos and video-sequences (eye-brow, and lip movements), gestures (hand and body movement) and physiological measures (temperature, humidity, heart rate, skin conductance, etc.). Some databases are collected accounting of several modalities simultaneously.

There is a considerable amount of image and audio emotional databases and a testimonial presence of physiological ones. Physiological measures of emotional states mainly refer to heart rates, skin temperature variations and electro-dermal activity. Such measurements always require the involved subject to wear a sensor which, no matter how comfortable it may be, can affect the physiological measurement.

It is worth mentioning on this respect the works of Kataoka et al. [12], Shusterman et al. [19] as well as Aubergé et al. [1] and Kim et al. [13] who implemented a 24-hour wearable ring or a wristwatch-type sensor to measure natural skin temperature (SKT) variations due to emotional stimuli.

The first to hypothesize that emotional feelings or stress may change the distribution of face temperature was Fumishiro [10] who used a thermal imager (the resolution was higher than 0.01ºC) to show that under emotional feeling the radiance temperature of eyes, noses and brows can vary in the range +/-0.2ºC. However, to date, there are no systematic studies linking face temperature and emotions. This work aims to scientifically test such a relationship by collecting a database of thermal and visible facial emotional expressions. The collected data will allow to assess if such changes can be considered an emotional feature and whether different emotions can be discriminated by different temperature values of the face or of regions of it.

To collect such data, emotions were induced though a carefully assessed experimental set-up (describe below) and the acquired database consisted of appropriately synchronized thermal and visible facial expressions. A selection of what were considered the most significantly emotional faces was also made using a custom developed Matlab software program. All the data, including those selected as best representatives of a given facial emotional expression, are available to the scientific community as part of the COST Action 2102 (http://cost2102.cs.stir.ac.uk/) activities. In addition, the present paper reports experimental data ascertaining the effects of emotions on memory word recognition tasks by measuring the individual recognition performance.

2 Database Design

The aim of this work was to define a database of emotional facial expressions which could be considered as “much spontaneous as possible”. The original idea was to identify video stimuli that could be used to elicit emotional states. Four emotions were selected among the six listed by Ekman [6] as basic emotions: *fear*, *happiness*,

sadness, and *disgust*. A *neutral* state was also considered[2], intended here as a state where no emotion is induced. This was done for the practical reason to separate series of facial video sequences recorded under a given induced emotion from another one as well as, to control the effects of an emotional stimulus on the other. *Surprise* and *anger* were not considered, due to the difficulty to elicit such emotional states through video stimuli. The definition of the spontaneous emotional facial expression database passed through three steps: 1) The identification of video stimuli to elicit the emotions under consideration – i.e. how the video-clips were selected and assessed; 2) The identification of a memory word recognition task, acting as a distractive task for restoring the subject's neutral state; 3) The acquisition protocol.

2.1 Identification of Video Stimuli with High Emotional Content

A total of 60 video-clips[3], 10 for each of the abovementioned emotional states were downloaded from YouTube (www.youtube.it) using the emotion labels as keyword. The original audio-track was kept. These stimuli were assessed by 20 naïve Italian subjects (9 males and 11 females) asked to watch the video-clips (randomly presented through a PPT Presentation) and label them by using the most appropriate of the 5 abovementioned emotional categories or any other emotional label. In addition, subjects were asked to rate the intensity of the portrayed emotion by using a Likert scale [14] varying from 1 (very weak) to 5 (very strong) through the intermediate values of 2 (weak), 3 (medium), and 4 (quite strong).

The result of this assessment was the identification of 5 video-clips for each emotion category (happy, sad, disgust, fear), plus 5 short neutral video-clips (30 sec.) separating an emotional video from another in the same emotional category, and 3 long neutral (2 minutes) video-clips separating sequences of different emotional category. This amounted to a total of 28 selected video-clips constrained to an average intensity rate value no lower than 3.

2.2 Identification of the Memory Word Recognition Task

In order to avoid overlaps among the induced emotional categories, a word memory recognition task was defined. In literature [15] such tasks are also called "recognition" tasks and consist of: a) A learning phase, where the subject memorizes a list of words (in our case 8 Italian words); b) A retention phase, where the subject is involved in an activity that has nothing to do with the task (in our case she/he was watching a sequence of 5 emotional video-clips belonging to the same emotional category inter-lived with short neutral stimuli); c) A re-enactment phase in which the subject is presented with a new list of words and she/he must provide a YES (if the word was already in the word list previously seen) or NOT (otherwise) answer .

---

[2] The authors of the present paper do have reservations on the existence of a *neutral* natural feeling but the discussion is out of the scope of the present work.

[3] The selected video-clip length varied from 30 to 140 sec. Longer stimuli were needed to induce sadness and/or to restore the neutral feeling in the subjects.

To this aim 8 word lists were created, 4 named Memory Lists (ML) and 4 named Recognition Lists (RL) each containing 8 Italian words. Both the word lists were shown on a computer screen. In each RLi there were 4 words already presented in the associated MLi , i:=1, .., 4. Before the induction of any of the 4 abovementioned emotional states, the subject was asked to read and memorizes the words in an MLi. Then, she/he was asked to watch a sequence of 5 emotional video-clips all belonging to the same emotional category. Finally, the RLi associated to the previously presented MLi was presented to the subject and she/he was asked to indicate on a paper grid, whether or not the words in the RLi list were already in the previously seen MLi one. A total of 48 almost equally frequent bi-syllabic, and three-syllabic Italian words were selected from the *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS*), www.alphalinguistica.sns.it. CoLFIS contains about 3.798.275 Italian words classified for frequency and complexity (monosyllabic, bi-syllabic, three-syllabic, etcetera). The words were randomly assigned to the 8 lists, taking care that each RLi contained 4 words already included in the corresponding MLi i=1, …4, the corresponding RLi i=1, …4. As an example the Ml1 and RL1 lists are reported in Table 1. The shared words are in bold.

Table 1. The ML1 and RL1 word lists. The shared words are in bold.

ML1 ITALIAN WORDS	ML1 TRANSLATION	RL1 ITALIAN WORDS	RL1 TRANSLATION
LUOGO	**PLACE**	**CAMPO**	**FIELD**
TITOLO	TITLE	PIANO	FLOOR
VALORE	**VALUE**	METRO	METRE
AZIONE	ACTION	**LUOGO**	**PLACE**
FUTURO	FUTURE	SERIE	SERIES
PASSO	**STEP**	**PASSO**	**STEP**
CAMPO	**FIELD**	**VALORE**	**VALUE**
SEGNO	SIGN	PEZZO	PIECE

2.3 Acquisition Procedure: The Experimental Set Up

The subject was invited to sit in front of a computer screen in order to perform the task which consisted of the following steps:

1. Read and memorize an MLi list in 30s;
2. Watch a set of 5 video-clips belonging to a given emotional category, each interleaved by a short neutral stimulus (N);
3. Read the RLi list associated to the previously seen MLi list;
4. Using a pencil and a YES or NOT answer signs the words in the RLi list seen in the MLi list;
5. Watch a Long Neutral (LN) stimulus;
6. Go back to step 2 until the end of the stimuli.

The stimuli presentation was randomized among the subjects according to the 4 different condition schemes reported in Table 2, where the letters indicate emotional categories, with S=sad, H=happy, F=fear, D=disgust. The facial expressions recorded from each subject were taken at 1 sec. sampling rate.

Table 2. Stimuli sequencing in each of the 4 identified CONDITIONS (A, B, C, and D). The letters indicate the emotional categories, with S=sad, H=happy, F=fear, D=disgust, N=Neutral. The number after the letter identifies the stimulus inside the category. For example, F3 indicates the third stimulus used for Fear. Note that the Neutral stimuli were always the same, but were associated randomly to the categories.

CONDITION A	**ML1 F1N1F2N2F3N3F4N4F5N5 RL1LN1 ML2 H1N1H2N2H3N3H4 N4 H5N5RL2LN2 ML3S1N1S2N2S3N3S4N4S5N5 RL3 LN3 ML4 D1N1D2 N2D3N3D4N4D5N5 RL4**
CONDITION B	**ML2** D1N1D2N2D3N3D4N4D5N5 **RL2LN2ML3** F1N1F2N2F3N3F4N4F5 N5**RL3LN3 ML4** H1N1H2N2H3N3H4N4H5N5 **RL4LN1ML1** S1N1S2 N2 S3N3S4N4S5N5 **RL1**
CONDITION C	**ML3** S1N1S2N2S3N3S4N4S5N5 **RL3LN3ML4** D1N1D2N2D3N3D4N4 D5N5 **RL4LN1ML1** F1N1F2N2F3N3F4N4F5N5 **RL1LN2 ML2** H1N1H2 N2H3N3H4N4H5N5 **RL2**
CONDITION D	**ML4** H1N1H2N2H3N3H4N4H5N5 **RL4LN1ML1**S1N1S2N2S3N3S4N4 S5N5**RL1LN2 ML2** D1N1D2N2D3N3D4N4D5N5**RL2LN3ML3** F1N1F2 N2F3N3F4N4F5N5 **RL3**

2.4 Hardware and Software Configuration

The acquisition system was equipped with two cameras each one connected to a separate laptop. A thermal camera TESTO 880-3 was connected to laptop SONY VGN-NS21Z with CPU Intel Core 2 Duo P8600 2.4GHz, 4GB RAM, MS Vista. This camera provided a 160 × 120 pixel resolution and the temperature range was set to 23-38°C (this camera provides a NETD < 0.1ºC. NETD is the Noise Equivalent Temperature Difference, which is a measurement of the sensitivity of a detector of thermal radiation). The visible camera was a Logitech Webcam C250 connected to laptop SAMSUNG NP R519 LED/T4200/2GB/250GB/SHARED/15.4/ VHP. This camera was set for a 640 × 480 pixel image resolution. Yawcam software, version 0.3.3, was used on both laptops, for the acquisition.

In both cases an image file (with a timestamp in the name) was acquired every 1 sec. in *.png format, in order to avoid huge memory space occupancy and consequently video compression. This is relevant especially for the thermal camera, sincethermal imagers do not provide as much resolution as webcams. Thus, it is important to keep the highest possible level of details without introducing a compression algorithm. The 1 sec. sampling rate was considered a good compromise between storage requirements and temporal-spatial resolution, since typical changes of muscular activities lasts for a few seconds [8].

2.5 Acquisition Scenario and Timing

The data collection was made in a quiet laboratory. Neither the acquisition computers, or the operators, or other people were visible to the participants. She/he watched the stimuli on a third laptop while wearing headphones to listen to the original video-clip audio-tracks, seated on a comfortable chair, with a black background and fluorescent room illumination, as illustrated in Figure 1. The acquisition took place from the 15th to the 19th of March 2010, between 9 a.m. and 6 p.m All the donors were Italian psychology undergraduate students, aged from 21 to 28 years. Such a population was deliberately chosen in order to reduce age and cultural background variability.

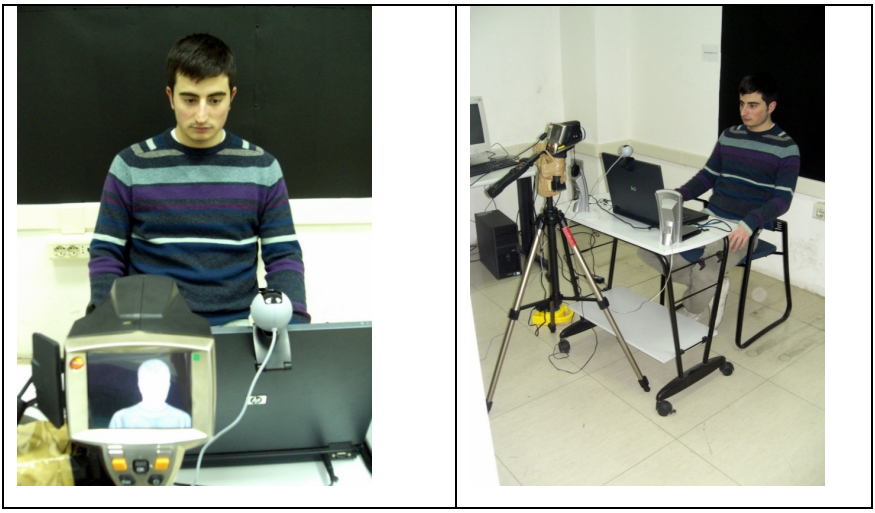


Fig. 1. Acquisition scenario

A consent form was filled and signed by each participant allowing the use of the collected data for scientific scopes. The acquisition timing for each subject is reported in Table 3

Table 3. Acquisition timing for each participant

Time elapsed	Tasks	Recording	Dialogue
5 minutes	Explanation of "what to do". Signature of the consent form	No	Yes
34s	Instructions on the computer screen	Yes	No
37.85 minutes	Data collection according to Table 2 recording schema	Yes	No
3 minutes	Subject comments and impressions	Yes	Yes

2.6 Database Description

More than 120.000 images for each camera (both the visual and thermal one) were collected during the experiment. Quantitative results obtained by human inspection are beyond the aims of this paper and may be tackled in future works.

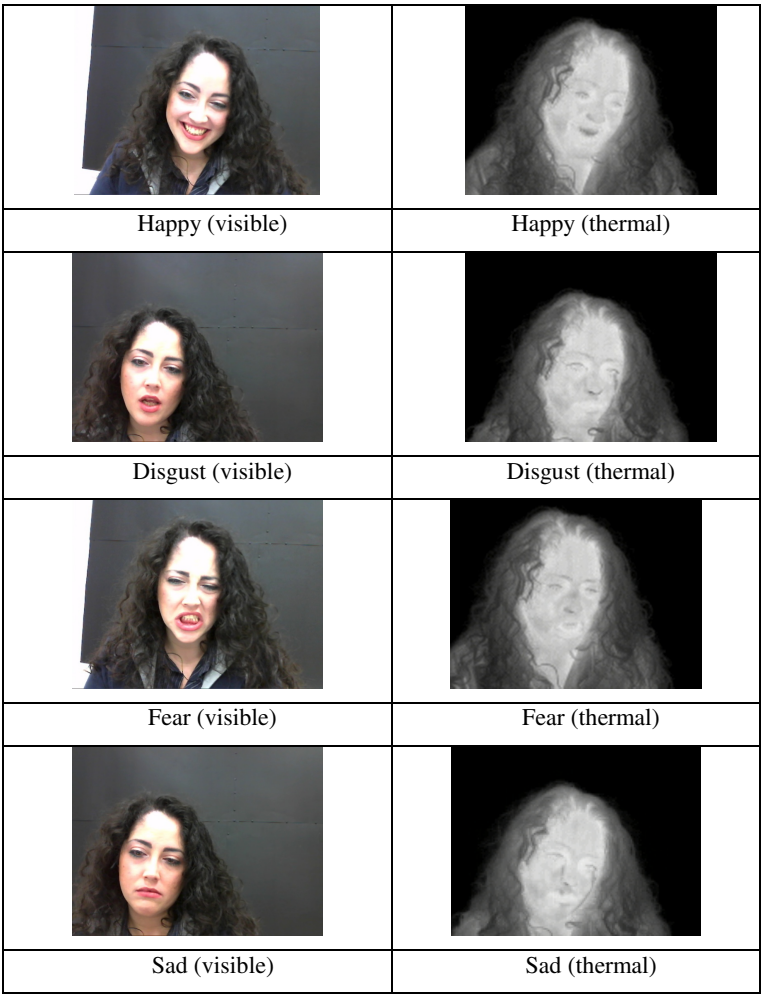


Fig. 2. Visible and thermal examples of emotional facial expressions

A snapshot in the visible and thermal domain of each induced emotional facial expression is displayed in Figure 2. It is worth noting that in the sad state the subject is crying and tears can clearly be seen in the thermal but not in the visible image.

2.7 Summary of the Main Characteristics

The collected database was named Italian Visible-Thermal Emotion (I.Vi.T.E.) database and will be freely distributed to the scientific community after the publication of this work. The main characteristics of the database are the following:

- Temperature range: 23-38 ºC;
- Size of database: 29.8GB (Thermal: 1.2GB, Visible: 28.6GB). Consists of one image per second in .png format;
- Total number of subjects: 49 Italian undergraduate students ranging from 22 to 28 years) watching sequences of highly emotional video-clips and listening to the original audio-tracks through headphones;
- Emotional categories under examination were: happiness, sadness, disgust, fear, and neutral;
- Thermal image resolution: 160 × 120 pixels;
- Visible image resolution: 640 × 480 pixels.
- Acquisition cameras: thermal (testo 880-3) and webcam (Logitech).

Using a custom Matlab software program the authors selected a total of 479 thermal and 479 visual images as the most significant facial emotional expressions elicited in the subjects. An example is displayed in Figure 3.

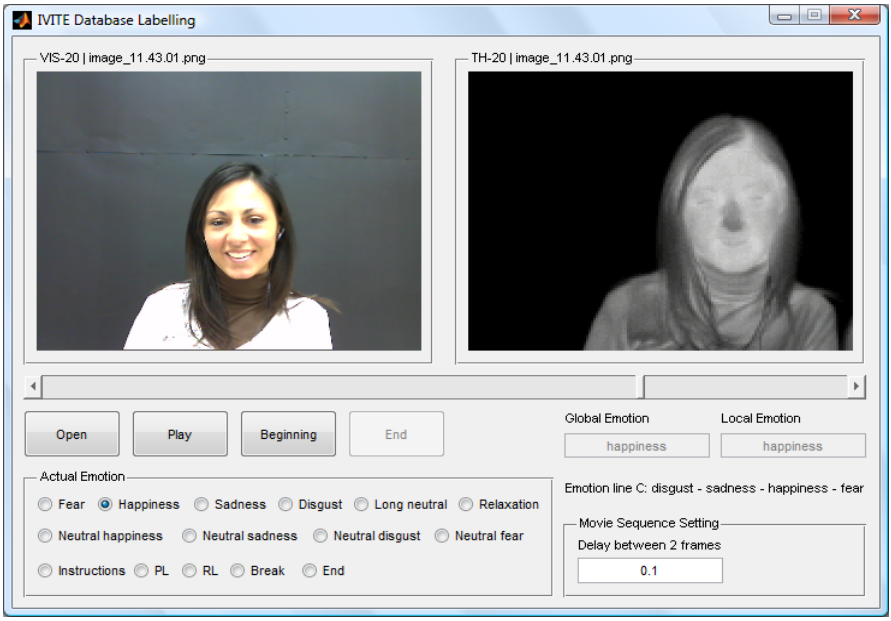


Fig. 3. A display of the software for manual labelling

This work was necessary in order to eliminate, amidst all the captured images, those which, according to a couple of expert judges, did not belong to the emotional categories selected for the experiment. The software is able to name each selected image, showing the type of camera used (thermal or normal), the number assigned to the participant (1 to 49), the timestamp of the collected image (expressed in minutes, sec., and msec.) and the temperature in Celsius degrees reported by the thermal camera.

3 Results on the Word Memory Recognition Task

The effects of the emotional states on the word memory recognition task were assessed considering the averaged error committed by each subject on the RL lists, after watching a given sequence of emotional stimuli, all belonging to the same emotional category.

Table 4. The number of words in the RLi lists wrongly listed by the subjects in each experimental condition

	Condition A	Condition B	Condition C	Condition D
Fear	18	16	5	12
Disgust	25	20	7	15
Happiness	13	18	24	14
Sadness	19	13	12	20

The original scores are reported in Table 4 for each of the four experimental conditions and for each emotion category. The numbers indicate how many words were wrongly listed in the RLi, i:=1,…4, lists by the subjects involved in a given experimental condition (A, B, C, D). Table 5 reports their transformation into z-scores with standard deviation σ equal to 5,47.

Table 5. Z-score transformation of the data reported in Table 4 with σ = 5,47

Z score	Condition A	Condition B	Condition C	Condition D
Fear	0,42	0,06	-1,95	-0,67
Disgust	1,70	0,79	-1,59	-0,12
Happiness	-0,49	0,42	1,52	-0,30
Sadness	0,60	-0,49	-0,67	0,78

As exposed in Tables 5 and 6, there were no significant differences in the subject's memory performance that could be attributed to a given induced emotional state or to a given experimental condition. None of the Z-scores falls outside the average score distribution in the real interval of [-2, +2]. Even the C condition, where both the best

and worst subjects' performance were gathered, does not show any significant deviation. The average word error in the word memory recognition task performed by the subjects on each emotional video sequence and for each of the 4 random elicited conditions is graphically displayed in Figure 4 and it varies in the real interval [0 2]. The average total error is illustrated in Figure 5. The data suggests that none of the induced emotional categories affects the word memory performance.

4 Results on the Thermal Data

The selected highly emotional faces, as reported in section 2.7, were manually tagged on 5 face regions (left part of left eye (LL), right part of left eye (RL), left part of right eye (LR), right part of right eye (RR), tip of nose (TN)) in order to measure possible changes in their temperature (with respect to the neutral state) when a given emotional state was induced.

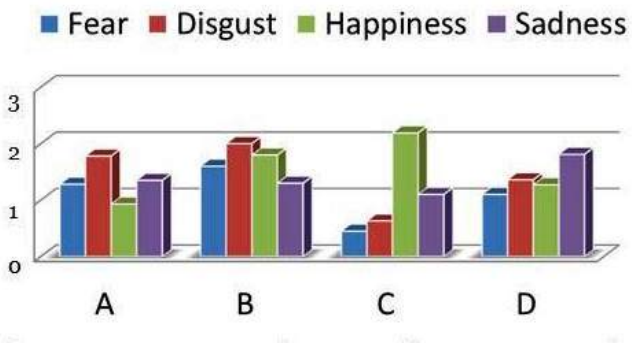


Fig. 4. Average word errors on the word memory recognition task distributed along the 4 experimental conditions

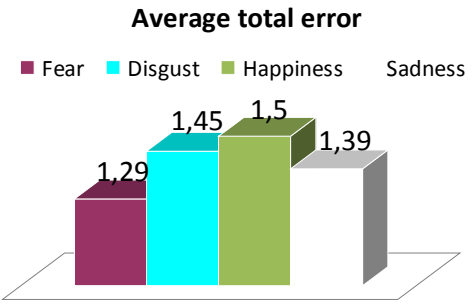


Fig. 5. Average total word errors on the word memory recognition task

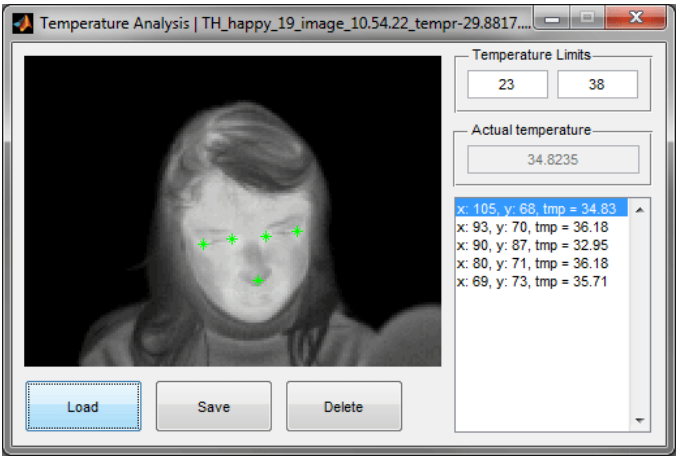


Fig. 6. Temperature analysis tool used for tagging the 5 face regions abovementioned

The custom Matlab software used for the tagging is illustrated in Figure 6. The temperature of these face regions was extracted from a 5x5 pixel matrix created around the selected points (as illustrated in Figure 6).

Table 6. Temperature data for the experimental condition A (females)

	Disgust						Fear					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
40	-0.72	-0.1	-0.34	-1.74	-0.16	-0.74	0.4	0.4	0.39	3.59	0.57	0.12
41	-0.56	-0.48	-0.08	-0.29	-0.66	0.29	0.28	-0.65	0.99	0.49	0.16	0.55
42	-0.28	0.28	0.21	-2.37	-0.04	-0.32	0.07	-1.08	-0.39	2.4	-0.51	-0.29
43	-1.14	0.1	-0.56	-5.52	0.24	-0.71	0.48	-0.24	-0.23	6.43	-0.33	0.24
44	1.12	-1.18	0.46	1.01	0.73	-0.48	-0.33	-0.51	-0.2	-0.14	-0.35	0.12
mean	-0.32	-0.28	-0.06	-1.78	0.02	-0.39	0.18	-0.41	0.11	2.56	-0.09	0.15
std	0.86	0.58	0.41	2.47	0.51	0.42	0.32	0.55	0.57	2.63	0.45	0.3
	Happiness						Sadness					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
40	-0.61	0.4	0.23	-0.11	0.41	-0.66	-0.62	-0.19	-0.43	-2.72	-0.98	-0.31
41	-0.05	0.03	-0.16	-0.29	-0.49	0.46	-0.5	0.12	0.49	1.27	0.08	0.29
42	-0.42	-1.66	0.18	1.48	0.15	-0.21	0.15	0.28	0.91	0.78	0.72	-0.21
43	2.35	0.05	0.04	5.67	0.71	1.11	0.04	0.69	0.75	2	0.24	-0.97
44	-1.11	-1.18	0.46	-0.37	-0.26	-0.48	-0.2	-0.68	-0.03	-0.71	0.81	-0.65
mean	0.03	-0.47	0.15	1.28	0.1	0.05	-0.23	0.05	0.34	0.12	0.17	-0.37
std	1.35	0.9	0.23	2.57	0.49	0.73	0.33	0.51	0.56	1.88	0.72	0.47

In addition, also the mean temperature of the whole face (WF) was considered. The measurements of the relative temperature changes (measured in Celsius degree relative changes with respect to the neutral state) are reported, as an exemplification and only for the experimental condition A, in Tables 6 and 7 for the females and males respectively. The gray columns indicate that 80% of the participants exhibited

Table 7. Temperature data for the experimental condition A (males)

	Disgust						Fear					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
46	2.28	-0.03	-0.36	-3.52	-0.22	-0.57	1.14	-0.37	1.45	0.56	0.66	-1.14
47	1.44	0.46	0.57	0.3	0.66	0.53	0.35	0.96	0.32	-0.88	0.74	1.13
48	-0.81	-1.45	-0.91	-6.07	-0.78	-2.81	0.09	0.14	0.77	7.08	0.69	0.86
mean	0.97	-0.34	-0.23	-3.1	-0.11	-0.95	0.53	0.24	0.85	2.26	0.7	0.28
std	1.6	0.99	0.74	3.21	0.73	1.71	0.55	0.67	0.57	4.24	0.04	1.24
	Happiness						Sadness					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
46	0.13	-0.03	0.79	-1.8	0.66	-0.97	0.18	-0.46	-0.53	-2.94	-0.58	-0.54
47	-0.56	-0.2	0.32	0.3	0.58	-0.07	1.06	-0.53	-0.48	-1.21	-0.55	-1.27
48	0.62	-0.4	0.6	3.69	0.69	-0.57	-1.67	-1.4	-0.46	-4.02	-0.46	-1.25
mean	0.07	-0.21	0.57	0.73	0.64	-0.54	-0.14	-0.8	-0.49	-2.73	-0.53	-1.02
std	0.59	0.18	0.23	2.77	0.06	0.45	1.39	0.53	0.03	1.42	0.06	0.42

in such face regions a temperature change with respect to the neutral state, measured before the emotion was induced. However, these changes randomly appear in different face regions when the experimental conditions change from A to B, C, D.

Table 8. Temperature data for the experimental condition B (females)

	Disgust						Fear					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
6	-0.42	1.04	0.22	-5.85	-0.44	-0.54	0.24	1.19	0.59	2.26	-0.49	-0.59
7	-0.8	1.79	-0.2	1.5	0.08	0.46	-1.37	1.14	0.56	2.21	0.65	-0.75
8	-0.86	-1.43	-0.81	-3.48	-1.08	-1.28	0.02	1.21	0.32	6.44	0.22	-0.11
11	-0.03	0.17	0.08	-0.93	-0.17	0.21	0.01	-0.68	-0.76	1.13	-0.59	-0.98
12	0.61	0.23	0.03	-1.59	-0.26	-0.63	-0.6	-0.92	-0.53	-0.74	-0.26	-0.63
14	-0.76	-0.13	-0.3	-1.7	-0.47	-0.37	-0.17	-0.05	0.2	-0.4	0.51	0.97
15	-0.02	0.24	-0.02	-4.12	-0.21	-0.23	0.13	0.09	0.37	-1.21	-0.18	-0.04
16	0.49	0.9	0.15	-1.29	0.06	-0.55	-1.02	-1.1	0.23	-1.18	-0.02	-0.55
17	-1.44	-0.77	-0.03	-1.23	-0.46	-0.7	-0.58	-0.29	-0.08	-0.91	0.11	-0.21
mean	-0.36	0.23	-0.1	-2.08	-0.33	-0.4	-0.37	0.07	0.1	0.84	-0.01	-0.32
std	0.68	0.96	0.31	2.13	0.35	0.51	0.56	0.92	0.47	2.52	0.42	0.58
	Happiness						Sadness					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
6	-0.02	0.22	0.11	5.98	0.06	0.11	-0.36	0.8	0.3	-4.32	-0.06	-0.28
7	-0.71	1.06	0.08	-1.46	0.73	-1.2	1	0.32	-1.13	-1.81	-1.01	-1.12
8	-0.53	1.21	0.24	5.11	0.14	0.14	-1.63	-0.19	0.08	-4.19	-0.11	-1.36
11	-0.21	-0.59	-0.42	2.37	-0.34	-0.21	-0.11	-0.68	-0.76	-0.51	0.25	1.58
12	-1.07	-0.51	-0.05	0.02	-0.34	-0.71	-0.07	0.23	0.11	-1.34	-0.26	-0.13
14	-0.31	0.03	0.03	0.73	0.59	0.64	-0.71	-0.91	-0.24	-1.52	0.04	-0.37
15	0.15	0.85	0.86	0.12	0.85	0.62	-0.98	-0.41	-0.12	-1.94	-0.18	-0.1
16	-0.17	0.34	0.39	0.79	0.82	1.77	-0.23	-0.6	-0.37	-0.72	-0.15	-0.08
17	-1.01	-1.07	0.25	-0.69	-0.77	-0.71	1.35	0.88	0.36	2.17	-0.05	-0.6
mean	-0.43	0.17	0.17	1.44	0.19	0.05	-0.19	-0.06	-0.2	-1.58	-0.17	-0.27
std	0.43	0.79	0.35	2.56	0.59	0.9	0.92	0.65	0.49	1.95	0.35	0.83

Table 9. Temperature data for the experimental condition B (males)

	Disgust						Fear					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
9	-0.69	-0.58	0.4	-1.68	0.33	0	-1.3	0.58	0.64	1.76	0	-0.67
10	0.44	-0.38	-0.08	-1.66	0.32	0.16	0.08	-1.13	-1.14	-3.14	-0.65	0
13	-0.5	-0.63	-0.56	-1.99	-0.64	-0.76	-0.38	-0.85	0.03	-0.85	0	0.23
mean	-0.25	-0.53	-0.08	-1.77	0	-0.2	-0.53	-0.47	-0.15	-0.74	-0.22	-0.15
std	0.61	0.13	0.48	0.19	0.56	0.5	0.7	0.92	0.91	2.45	0.37	0.47
	Happiness						Sadness					
P. ID	WF	LL	RL	TN	LR	RR	WF	LL	RL	TN	LR	RR
9	-0.74	-0.58	0.7	1.94	-0.11	-1.68	0.22	-0.58	-0.16	-0.88	-0.24	0.67
10	0.2	1.36	0	-1.74	0.49	0.16	-0.43	-1.05	-0.16	-3.01	-0.53	0.04
13	0.48	0.3	0.03	-2.39	-0.16	0.23	-0.61	0.3	0.03	-1.03	0	0.4
mean	-0.02	0.36	0.24	-0.73	0.07	-0.43	-0.27	-0.44	-0.1	-1.64	-0.26	0.37
std	0.64	0.97	0.39	2.34	0.36	1.09	0.43	0.68	0.11	1.19	0.26	0.32

For example, the temperature changes for the same emotional category in the experimental condition B (see Tables 8 and 9 for females and male respectively) do not follow the same pattern observed for the experimental condition A (see Tables 6 and 7).

Therefore, it seems that with this temperature resolution and in the defined experimental conditions, emotional states do not significantly change the temperature of the face or regions of it. An increased temporal-spatial resolution of the thermal camera to identify appreciable temperature changes would be necessary.

5 Conclusions

This paper reports on a collection of naturalistic thermal and visible induced facial emotional expressions providing details on the experimental set-up, the acquisition scenario, the eliciting stimuli and the data. Facial emotional expression recognition through visible images has occupied a great deal of research, while thermal images have not yet been considered. Given that thermal images have the good property of not being affected by illumination and shadows, they can be, to a certain extend, more useful than the visible ones to determine distinctive facial emotional features.

In addition, this work reports data obtained through a pilot experiment, showing no effects of emotional states for a defined word memory recognition task. It could be argued that the proposed word memory recognition paradigm (memory task) proved to be ineffective by the emotional interference, compared to the recall paradigm proposed by Dougherty and Rauch [5]. However, this rises several open questions on the intervention of emotional states on memory performance. Further investigations are needed to assess which are, and to what extent cognitive and memory tasks are affected by emotional states. Some questions. Which emotional state will produce an improvement or a deterioration of the cognitive and memory performance? Does the

feeling experienced by the subject in the learning or the retention phase play a role in the accuracy of the recognition? For a better memory performance, is the emotional feeling state at the time of the encoding more important than the one experienced during the retention of the mnemonic material? Literature suggests the importance of both [16-18]. However, more data are needed. Finally, what are the effects of the sequencing? Does it produce a bias in the learning and retention phase?

Finally, it was shown that an increased temporal-spatial resolution of the thermal camera would be necessary to observe appreciable temperature changes in the face or regions of it.

References

1. Aubergé, V., Audibert, V., Rilliard, A.: Why and how to control the authentic emotional speech corpora. In: Proc. of 8th European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland, pp. 185–188 (2003)
2. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to find trouble in communication. Speech Communication 40, 117–143 (2003)
3. Cheon, Y., Kim, D.: Natural facial expression recognition using differential-AAM and manifold learning. Pattern Recognition 42, 1340–1350 (2009)
4. Cowie, R., Douglas-Cowie, E., Cox, C.: Beyond emotion archetypes: databases for emotion modelling using neural networks. Neural Networks 18 (2005)
5. Dougherty, D.D., Rauch, S.L.: Brain correlates of antidepressant treatment outcome from neuroimaging studies in depression. Psychiatric Clinics of North America 30(1), 91–103 (2007)
6. Ekman, P.: Universals and Cultural Differences in Facial Expressions of Emotions. In: Cole, J. (ed.) Nebraska Symposium on Motivation, pp. 207–283. University of Nebraska Press (1972)
7. Ekman, P., Friesen, W.: Facial Action Coding System (FACS): A technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
8. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition 36, 259–275 (2003)
9. Faundez-Zanuy, M., Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Multimodal biometric databases: An overview. IEEE Aerospace and Electronic Systems Magazine 21(8), 29–37 (2006)
10. Fumihiro S., Ma L.: Measurement of face temperature distribution by thermal imager. Papers of Technical Meeting on Light Application and Visual Science, IEE Japan, vol. LAV-99 (1-6), pp. 29-33 (1999)
11. Izard, C.E.: Basic emotions, relations among emotions. and emotion-cognition relations. Psychological Review 99, 561–565 (1992)
12. Kataoka, H., Kano, H., Yoshida, H., Yasuda, M., Osumi, M.: Development of a skin temperature measuring system for non-contact stress evaluation. In: IEEE Ann. Conf. Engineering Medicine Biology Society, Hong Kong, pp. 940–943 (1998)
13. Kim, K.H., Bang, S.W., King, S.R.: Emotion recognition system using short-term monitoring of physiological signals. Medical and Biological Engineering and Computing, 419–427 (2004)
14. Likert, R.: A technique for the measurement of attitudes, vol. 22, p. 40. Archives of Psychology, New York (1932)

15. Roediger, H.L., Wheeler, M.A., Rajaram, S.: Remembering, knowing, and reconstructing the past. In: Medin, D.L. (ed.) The Psychology of Learning and Motivation: Advances in Research and Theory, pp. 97–134. Academic Press, San Diego (1993)
16. Schacter, D.L.: Searching for Memory. The Brain, the Mind and the Past. Basic Books, New York (1996)
17. Schacter, D.L., Singer, J.E.: Cognitive, social and physiological determinants of emotional state. Psychological Review 69, 379–399 (1962)
18. Schacter, D.L.: Memory distortion: History and current status. In: Schacter, D.L., et al. (eds.) Memory Distortion. Harvard University Press, Cambridge (in press)
19. Shusterman, V., Barnea, O.: Analysis of skin-temperature variability compared to variability of blood pressure and heart rate. In: IEEE Ann. Conf. Engineering Medicine Biology Society, Montreal, pp. 1027–1028 (2005)

Prosody Modelling for TTS Systems Using Statistical Methods

Zdeněk Chaloupka and Petr Horák

Institute of Photonics and Electronics, Academy of Sciences of the Czech Republic
{chaloupka,horak}@ufe.cz

Abstract. The main drawback of older methods of prosody modelling is the monotony of the output, which is perceived as uncomfortable by the users, especially when listening to longer passages. The present paper proposes a prosodic generator designed to increase the variability of synthesized speech in reading devices for the blind. The method used is based on text segmentation into several prosodic patterns by means of vector quantisation and the subsequent training of corresponding HMMs (Hidden Markov Models) on F0 parameters. The path through the model's states is then used to generate sentence prosody. We also tried to utilize morphological information in order to increase prosody naturalness. The evaluation of the quality of the proposed prosodic generators was carried out by means of listening tests.

Keywords: Text-To-Speech, HMM, prosody, morphological information.

1 Introduction

The degree of naturalness of synthesized speech is largely determined by the F0 profile of the utterance. However, the prosody of each language varies considerably depending on the speaker's characteristics as well as the content of the message. Unit selection based TTS systems may use source databases large enough to avoid prosodic „stiffness" and monotony in smaller applications, but the use of such systems is limited by database size. Common TTS systems, mostly using data from a small inventory, suffer from low quality of generated prosody due mainly to the algorithm used for prosody modelling (rule-based prosody, neural networks etc.). In everyday use of the application, e.g. by the blind listening to longer texts, the monotonous prosodic form of the TTS output is a substantial drawback which the users complain about.

In the articles [1, 2, 3], different methods of prosody generation are mentioned, the following being the most frequent:

- algorithms containing prosodic rules of the language in question (commonly formulated in terms of regular expressions)
- learning algorithms like neural nets, HMMs or FSMs
- hybrid processing techniques [3]
- post-processing of the aforementioned methods by means of different filters etc.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 174–183, 2012.

In this paper we shall discuss the development of an algorithm which is able to generate sufficiently variable prosody from a large database (over 5 000 sentences). The training methods, which are almost completely automatic, permit the treatment of large sets of data.

The paper is organized as follows. Section 2 summarizes the prosody generation background and presents motivation for specific parameters selection. The HMM training phase setup and inverse problem (prosody generation) are solved in the subsection 2.1. The Section 3 focuses on the listening tests. Finally, we conclude outcomes in the last Section.

2 Prosody Generation

Voice melody (as a component of prosody) is the main factor which allows to formulate the content of the message (e.g. by differentiating declaratives from questions or simple sentences from complex sentences). Consequently, the prosody of the utterance is determined especially by (1) sentence type and (2) its structure. Also, the length of the whole sentence (and its parts), the context and naturally, the personal characteristics of the speaker also have great influence on the sentence prosody. Since the context of the utterance is hard to determine unless sophisticated methods are used, we shall stick to points 1 and 2 (as mentioned above) in formulating the algorithm.

The input data available for the algorithm training are the following: read studio recordings of J. Heller's "God knows" (16 kHz, 16 bits), and the corresponding text in electronic format, obtained from the printed version by OCR. The reading speed was below the normal speed rates. It ranged between 150 and 200 Words Per Minute (WPM). The speed was lower deliberately, because the emphasis was put on the correct prosody output. The audio recording was aligned with the text and segmented into shorter sentences so as to make the automatic alignment of sound and word boundaries easier (for more details, see [4]) – this is actually the only stage of the algorithm training which requires manual work (and even this could be fully automated if sophisticated recognisers were used).

As mentioned in the first paragraph of this chapter, the prosody of a sentence is determined mainly by its type and its structure. It is also evident that the prosody of individual sentences occurring in a sequence will be different, because the speaker will not use the same intonation (two subsequent declarative sentences will never be read with a fully identical intonation). Furthermore, it is clear that the intonation will be different in the middle of a paragraph and at the end of it, where the sentence is followed by a longer pause. Another crucial aspect of prosody is the syllabic structure of the sentence, since syllables influence rhythm, and, subsequently, melody [4, 5].

It is the last word in the sentence that bears a significant part of the prosody, especially in the case of interrogative sentences or sentence tags. Therefore, we have to extract information about the prosody from the very last word of prosodic phrase.

Based on aspects mentioned above, we defined a set of parameters determining the prosodic type of the phrase in question (portion of the sentence separated by conjunctions or commas). Note that prosodic phrases were extracted using rule-based algorithm of the Epos system [12]. The parameters follow:

1. sentence type (declarative, imperative, interrogative, plus the semicolon and the dash as a specific category); sentences in quotation marks could be added as well (as their melody is more coloured in narratives)
2. number of sounds in %, i.e. the ratio between the number of sounds in the phrase and in the whole sentence
3. number of syllables in %, i.e. the ratio between the number of syllables in the phrase and in the whole sentence
4. number of words in %, i.e. the ratio between the number of words in the phrase and in the whole sentence
5. type of the pre-preceding sentence (see the first parameter), as well as the number of phrases contained in this sentence
6. type of the preceding sentence – see point 5
7. type of the following sentence – see point 5
8. type of the final sound (vowel, consonant)
9. number of vowels in the last word (in %) – ratio between the number of vowels (voiced sounds bear intonation) in the last word of the phrase and the number of all sound this word

These parameters were arranged in a set of vectors which served as the input for a simple vector quantifier (VQ), measuring the Euclidean distance from the centroids [6]. The VQ algorithm was set to find out 16 different centroids. That means it outputs a codebook containing the centroids of 16 prosodic patterns (differentiated by means of 9 input parameters mentioned above).

It should be noted at this point that the number of prosodic patterns may (and should) be extended. Also, more codebooks may be created, depending on the parameter (1). That means for each different parameter (1) there should be a different codebook. In this way we can achieve a greater prosodic variability while preserving the simplicity of the algorithm. The next logical step consists in assigning F0 characteristics to prosodic patterns of the codebook. A pre-requisite for this is the choice of a training algorithm which would best fit the character of the task.

Given the amount of data, the training of prosodic parameters (i.e. F0 values) should be based on an algorithm using statistical processing, i.e. including a learning (training) stage. The fact that F0 estimation is done automatically (see next paragraph), without manual correction – and thus with errors – also points towards the use of a statistical algorithm. In the light of these arguments, the authors of the present article consider that the best choice is the use of an HMM algorithm (HTK software package [14]), of which they have experience. Obviously, any algorithm complying with the parameters above can be used at this stage.

2.1 HMM Based Prosody Generator

The input of the HMM algorithm is represented by F0 values, which are to be extracted from the audio recordings. This was realised by means of the WaveSurfer software, which proved very accurate [7], and, moreover, permits batch processing of large amounts of data. The output F0 data (sampled every 10 ms) were processed so as to eliminate zero values of F0, which convey no prosodic information (corresponding to unvoiced sounds or intervals of silence). It is also possible to remove values which are extremely low (< 50 Hz) or extremely high for the given speaker (> 200 Hz), as these undoubtedly correspond to wrong estimations of the algorithm. However, since we use statistical methods, we did not consider this manipulation to be necessary. Next, the number of F0 samples must be rounded to a multiple of the number of samples in one HMM (as defined by the properties of the HTK software) – see the information about HMM type given below. The HMM prototype setting [14] is as follows:

- number of states: 7
- number of mixtures: 2
- matrix of transition states:

```
0.0 1.0 0.0 0.0 0.0 0.0 0.0
0.0 0.5 0.3 0.2 0.0 0.0 0.0
0.0 0.0 0.5 0.3 0.0 0.2 0.0
0.0 0.0 0.0 0.6 0.3 0.1 0.0
0.0 0.0 0.0 0.0 0.6 0.3 0.1
0.0 0.0 0.0 0.0 0.0 0.6 0.4
0.0 0.0 0.0 0.0 0.0 0.0 0.0
```

- data type: F0, *d*F0, *dd*F0 (where *d* stands for derivation), i.e. derivation and acceleration coefficients
- the number of samples per model is 5, and the corresponding number of parameters per model is 15 (5 x 3 samples – F0 + *d*F0 + *dd*F0)

It should be noted that every prosodic pattern mentioned in Section 2 is associated with one HMM. That means, each HM model describes one of the 16 prosodic patterns of the VQ codebook. These HMMs can be initialised in two ways [14]: by means of a free initialisation ("flat start" under HTK), or by means of a discrete initialisation of individual models. We used the latter method, which is more accurate, but requires the knowledge of model distribution in training data. Therefore, the time alignment of phrases in the given recording must be specified. To fulfil this condition, we first aligned the text with the recording using a classical speech recogniser (forced alignment under HTK). Next we converted word/sentence boundaries into prosodic pattern boundaries as the prosodic patterns are bound to specific parts of the text (i.e. phrase boundaries are before some conjunctions, sentence end etc. – see Fig. 1.). Once the text time alignment is known, subsequently, we obtain the corresponding

phrase boundaries in F0 values. It should be noted that system Epos, which was used as a text pre-processor, marked the prosodic phrase boundaries in text.

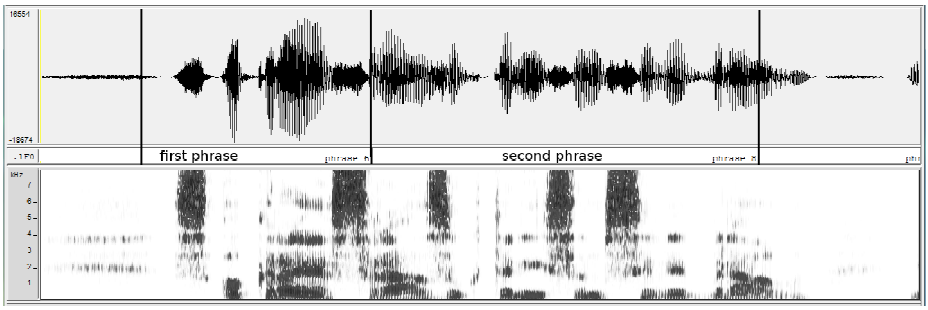


Fig.1. The picture shows speech amplitude (upper box – with prosodic phrase boundaries), text label of phrase (mid box) and spectrum (lower box) of the sentence

These time indications (phrase boundaries) are used for model initialization under HTK. The training phase was repeated several times using the whole set of data. The training of the models was then stopped despite the fact that we were still far from the algorithm convergence (the amount of time requested for reaching the convergence would be too much). The models had been trained to a certain degree anyway, and we could test their quality. However, concerning the training phase there is still room for experiments.

Although the common scheme is to divide the data to training and testing sets, we decided to use all the data for training. The main reason is that we didn't want to evaluate the quality of the HMMs by means of recognition. We preferred to use a listening test of the whole prosody generator as such, because it was supposed to provide us a better feedback. Before any evaluation took place, we first had to provide automatic generation of model states (i.e. prosodic values) in the TTS system, which turned out to be the hardest part of the task.

After a series of unsuccessful tests using different HMMs to generate model states observations from text-associated data (HM model based on individual sounds, their position in the sentence etc.), we finally used a much simpler method to obtain sequence of model state observations. The structure of the algorithm is expressed by the following formulae:

$$\text{let } b_k = a_{ij}, \text{ only for } i = j, \text{ where } a_{ij} \text{ is transition matrix}$$

$$0 \leq b_k < 1, k = 0..N, \text{ where } N \text{ is the number of the HMM states}$$

$$T = \sum_{m=0}^{P} p_m, \text{ where } p_m \text{ is time length of the } m-th \text{ phone}$$

$$\text{and } P \text{ is the number of the phonemes of prosodic phrase}$$

$$t_k = b_k * \frac{T}{\sum b_k} \tag{1}$$

$$\sum b_k t_k = T \tag{2}$$

$$b_k t_k < 30ms \Rightarrow b_k = 0 \tag{3}$$

From the definition of b_k one can assume that we are using only non-transition states of the HMM. There is N states of the HMM for one prosodic phrase (that is based on HMM definition). We were using 7 states for the HM model, that is 5 real states as the start and end states in HTK are virtual. Time T is computed as a sum of the current phones times in the prosodic phrase. Now the iteration through (1-3) takes place. For each HMM state we computed its duration using formulae (1). The t_k sequence contains information about the time for which the HMM remains in its *k-th* state. The total duration of all states must be equal to the time length T of the prosodic phrase (i.e. Eq. 2). If a HMM state duration is less than 30ms we skip such state (Eq. 3) as it wouldn't be audible anyway. The cycle is then iterated till all three conditions are satisfied. It should be noted here that the speed of the synthesized text is (for testing purposes) as close to the original as possible, so the WPM parameter is below 200. The selection of mixture for the given *i-th* model state is performed randomly, according to the following criterion:

$$\arg\max(w_{i,l} * r_{i,l}) \text{, where } l = 1..M, \text{ and } M \sim \text{number of mixtures}$$

$$w_{i,l} \text{ is } i\text{-}th \text{ model mixture weight, } r_{i,l} \text{ is randomly generated number}$$

The HMM-based prosody prediction is implemented in the Epos TTS system, which has been developed in the Institute of Photonics and Electronics (Academy of Sciences of the Czech Republic). The source code is written in C++, as is the case for the entire system. The process of speech generation is displayed in the following figure.

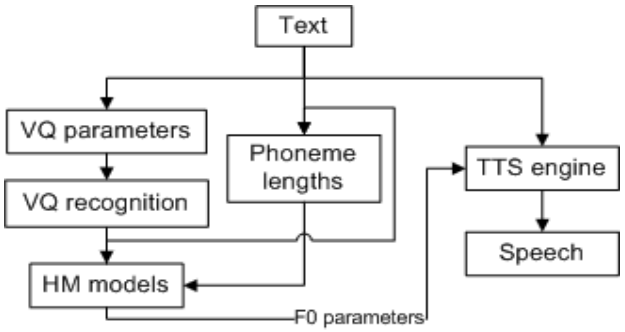


Fig.2. Scheme of prosody generation inside the Epos system

2.2 Lexico-Syntactical Prosody Generator

So far, we have systematically omitted morphological (or, more precisely, lexico-syntactic) information about the sentence in our experiments concerning prosody

generation. For the sake of simplicity, we shall use the term "morphological information", even if it is not fully adequate. The present chapter is concerned with basic integration of morphological information in the process of prosody generation.

The morphological information was retrieved by means of the TECTO MT software package, developed in the Institute of Formal and Applied Linguistics [9, 13]. This set of scripts provides the input text with morphological tags. The method of text transcription is described in [8, 9, 13]; we shall not discuss it in more detail in the present paper.

At this stage, each word in the sentence has been provided with a morphological tag. To simplify the procedure and reduce the number of combinations of morphological entities in the different parts of the sentence, let us assume that the most relevant part in terms of sentence prosody is the final word of the prosodic phrase (separated by a clause conjunction). The words at the end of each prosodic phrase will be classified into "morphological patterns" according to the following criteria: The first division is based on sentence type (declarative, interrogative, imperative), and the second division on morphological tags. The authors of the TectoMT software use the following 12 lexico-syntactic types:

- adjective, numeral, adverb, interjection, conjunction, noun, pronoun, verb, preposition, particle, unknown, punctuation

Since shorter words necessarily contain a smaller number of prosodic points, we classify the words into three size categories (short, middle-sized or excessively long). These rules account for a set of 3 x 3 x 12 different patterns.

Automatic prosody generation based on morphological tags requires the training of morphological patterns on the recorded database (more than 5,000 sentences). The training algorithm is based on the similarity (correlation) between the model and the trained data. The model is updated preferentially by data which are more similar to it.

3 Listening Tests and Results

The prosody generators were assessed by means of listening tests, in which generated sentences were rated along with real-speech recordings and with sentences produced by older (rule-based) version of the prosodic module. We also tested a combination of the two proposed methods, HMM and morphological tag based (MTG). Since the assessment of prosodic quality by comparing synthesised text with real speech is biased due to the imperfection of the synthesis as such, we re-synthesised the assessed text with the prosodic parameters extracted from real-speech recordings. The test was performed by fifteen listeners (for twelve different sentences), the listeners were told to evaluate the quality of the synthesized speech strictly from the prosodic point of view (i.e. questions with rising intonation in the end etc.). The results of the listening test are given in the following Table.

Table 1. The table shows results of the listening test for different types of prosody generation algorithm. The marks respond to this verbal assessment: 1 - natural, 2 - rather natural, 3 - rather unnatural, 4 – unnatural.

Type of synthesis	Mean mark	Variance of the marks
Original	2.49	1.19
Rule-based	2.21	0.81
HMM	2.75	0.70
HMM + MTG	2.82	0.63
MTG	2.60	0.80

The mean and variance marks can be somewhat misleading, therefore we adopt the standard paired t-test hypothesis test that assesses whether the means of two groups are statistically different from each other [10]. The t-test assumes that the distribution of the data is normal. We used Jarque-Bera test to determine the data normality [11].

Table 2. The table shows the Jarque-Bera test and pair t-test p-values at the 5% significance level. Diagonal - Jarque-Bera test p-values (italic). Rejected null hypothesis (bold typed).

Synthesis	Original	Rule-based	HMM	HMM+MTG	MTG
Original	***0.0156***	0.0190	0.0220	$3.30 \cdot 10^{-3}$	**0.3449**
Rule-based		*0.0629*	$2.48 \cdot 10^{-7}$	$3.37 \cdot 10^{-9}$	$2.79 \cdot 10^{-4}$
HMM			*0.1319*	**0.4716**	**0.1357**
HMM+MTG				*0.1415*	0.0267
MTG					*0.1159*

The results shown in Table 2 imply that the listening marks of original synthesis probably do not have the normal distribution required for t-test. So the t-test hypothesis results concerning the original synthesis should be thought of carefully. Also there is no significant difference (based on rejection of the t-test null hypothesis) between the following pairs of synthesis <original, MTG>, <HMM, HMM+MTG>, <HMM, MTG>. Hence, from statistical point of view, there is a little difference between them.

We also wanted the listeners to sort the speech recordings according to which of the prosody style is more likely less tiring after one hour of listening. The test was performed on a longer text (one minute of continuous speech). The results are given in the following Table.

Table 3. The algorithm priority for different persons

Person	Algorithm preference (the last is least preffered)			
1	Rule-based	HMM	HMM+MTG	MTG
2	Rule-based	MTG	HMM	HMM+MTG
3	HMM	HMM+MTG	MTG	Rule-based
4	HMM+MTG	HMM	MTG	Rule-based
5	HMM+MTG	MTG	HMM	Rule-based
6	HMM+MTG	HMM	MTG	Rule-based

4 Conclusion

As you can clearly see from Table 1 and 2 the classic old-fashioned rule based algorithm still seems as the most precise prosody generator in terms of Czech language intonation rules. Even the re-synthesized prosody sounded less natural in comparison to rule-based algorithm for most of the listeners. This could be explained by inconsistency in the real speech prosody generation (that stands for our training data as well), where the speaker's intonation do not always conform to Czech prosody rules as the speaker tries to emphasizes some part of the speech for example.

The second test (Table 3) strengthens our belief that the variability of prosody naturalness is perceived really differently. In this test, the listeners were told to sort the algorithms according to tediosity of the speech. Two listeners preferred the rule-based algorithm while the other gave preference to the HMM (1) and HMM+MTG (3) algorithms.

Following the purpose of the Epos software the only relevant criterion will be the real feedback from users with visual impairment (which is to be tested in future). Anyway, there is still room for improvement of the proposed algorithms, mainly concerning the initial setup of the prosodic phrases, extracted features and HM model training.

Acknowledgement. This research was realised with the support of the GA ČR 102/09/0989 grant project.

References

1. Rajeswari, K.C., Uma, M.P.: Prosody Modeling Techniques for Text-to-Speech Synthesis Systems – A Survey. International Journal of Computer Applications 39(16), 8–11 (2012)
2. Malfrère, F., Dutoit, T., Mertens, P.: Automatic Prosody Generation Using Suprasegmental Unit Selection. In: Proc. ESCA Workshop on Speech Synthesis, pp. 323–328 (1998)

3. Bellur, A., Narayan, K.B., Raghava, K.K., Murthy, H.A.: Prosody modeling for syllable based concatenative speech synthesis of Hindi and Tamil. In: National Conference on Communications, pp. 28–30 (2011)
4. Chaloupka, Z., Uhlíř, J.: Speech Defect Analysis Using Hidden Markov Models. Radioengineering (2007)
5. Hardcastle, W.J., Laver, J., Gibbon, F.E.: The Handbook of Phonetic Sciences (2009) ISBN 978-1-4051-4590-9
6. Deza, M.M., Deza, E.: Dictionary of distances. Elsevier (2006) ISBN-13: 978-0-444-52087-6
7. Bořil, H.: Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora, Ph.D. dissertation, Czech Technical University in Prague, Czech Republic (2008)
8. Hajič, J.: Complex Corpus Annotation: The Prague Dependency Treebank. Jazykovedný ústav Ľ. Štúra, SAV, Bratislava, Slovakia (2004)
9. Žabokrtský, Z., Ptáček, J., Pajas, P.: TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In: Proceedings of WMT (2008)
10. Sokal, R.R., Rohlf, F.J.: Biometry: The principles and practice of statistics in biological research, 3rd edn. W.H. Freeman, New York (1995)
11. D'Agostino, R.B.: Tests for the Normal Distribution. In: D'Agostino, R.B., Stephens, M.A. (eds.) Goodness-of-Fit Techniques. Marcel Dekker, New York (1986) ISBN 0-8247-7487-6
12. Epos system, `http://epos.ufe.cz`
13. Žabokrtský, Z., Bojar, O.: TectomMT - Developer's Guide, `http://ufal.mff.cuni.cz/tectomt/guide/guidelines.html`
14. HTK software, Ver. 3.2.1., `http://htk.eng.cam.ac.uk`

Modeling the Effect of Motion at Encoding and Retrieval for Same and Other Race Face Recognition

Hui Fang, Nicholas Costen, Natalie Butcher, and Karen Lander

Manchester Metropolitan University
School of Computing, Mathematics and Digital Technology,
Manchester, U.K.
n.costen@mmu.ac.uk

Abstract. We assess the role of motion when encoding and recognizing unfamiliar faces, using a recognition memory paradigm. This reveals a facilitative role for non-rigid motion when learning unfamiliar same and other-race faces, and indicate that it is more important that the face is learned, rather than recognized, in motion. A computational study of the faces using Appearance Models of facial variation, shows that this lack a motion effect at recognition was reproduced by a norm-based encoding of faces, with the selection of features based on distance from the norm.

1 Introduction

The two dominant theories of the effect of movement on face recognition are the supplemental information hypothesis and the representation enhancement hypothesis [1]. For familiar faces, the first suggests a person's characteristic facial motion is integrated into the face's representation. The second suggests that facial motion facilitates the perception of the three-dimensional structure of the face. As this is not dependent on previous experience with an individual face, it may be important in understanding how motion aids recognition and learning of faces. This second, facilitative, role of motion for unfamiliar faces may be a product of the construction of more robust mental representations at encoding.

At the same time, it is known that other-race faces are recognized with less proficiency than same-race faces [2]. There is no universally accepted agreement on the mechanisms responsible for the other race effect [3]. Furthermore, there has, to date, been no such investigation of the effect motion has on the recognition of other-race faces despite the known variation in the type of facial motions exhibited across different races [4].

If a lack of expertise in processing other-race faces explains the 'other-race effect', then this may include individuating motion information exhibited by other-race faces. However, familiarization with other-race faces reduces the other-race effect and increase levels of configural and holistic processing [5]. If viewing a moving face enables the creation of more robust, descriptive face representations, enhanced encoding of other-race faces may lead to motion being seen to be

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 184–190, 2012.

beneficial to the recognition of unfamiliar other-race faces as it is for same-race faces [6].

It is also necessary to consider the representation of the faces. A typical model which has received much attention in recent years [7, 8] is a norm-based description of face space[9, 10]. This codes faces as deviations from a central exemplar (typically the mean of the person's facial experience) with respect to weighting on a large set of parameters. There is evidence that these weightings are normalized by their variance [8]. Other-race faces may be represented in the same space, or in a smaller, specialized, one. The interaction between such norm-based face models and representational enhancement is unclear; this study seeks to investigate it, with particular attention to the issue of the consistency of the parameters on which the faces are encoded.

2 Psychological Data

This paper reports modeling of psychological results collected by the authors [11]. This used a recognition paradigm (faces were learned and then individually displayed amongst distractors), with four experimental conditions: static learning with static recognition, static learning with moving recognition, moving learning with moving recognition and moving learning with static recognition. In addition, parallel sets of British Caucasian and Japanese faces tested whether learning other-race faces in motion increases recognition.The sequences showed a mixture of rigid and plastic head motion, captured whilst the models spoke specified, short, sentences.

Assessed via Receiver Operating Characteristic (ROC) and reaction time, a significant main effect of race was found, indicating that participants were more sensitive when recognizing faces of the same race to their own (British Caucasian) than faces from another race (Japanese). There was also a significant main effect of learning presentation style; learning using motion gave accurate recognition than learning with static images. However, there was no significant difference between recognizing the target faces from static or moving images. There were no interaction effects.

3 Computational Analysis

A statistical model was built to simulate aspects of the human cognitive facial recognition system. Although this is primarily a computational model [12], the processing sequence has some similarities to human cognition. First, a set of consistent facial feature points are located; this provides configural information [13]. Second, a statistical facial model including a mixture of races (a majority of Caucasian faces and minority of Japanese faces) is built. Thirdly, noise is added to the feature-points located on the static faces to simulate the effects of structural enhancement [1]. Fourthly, distinctiveness levels are set to simulate human sensitivities to facial variations. Finally, statistical features which have significant variations are recorded and used to identify faces [9].

3.1 Group-Wise Registration

As the most important pre-processing step in facial recognition, salient feature point localization not only aligns facial features between different faces but also tracks the dynamic changes in a sequence. Our focus is on investigating the effects of motion on faces, requiring a good alignment and salient feature tracking method to capture the non-rigid deformations. The Group-wise Registration Framework [14] tracks salient facial features robustly, iteratively updating a model of the face currently under analysis to give dense correspondence between the images. A set of 68 salient points are then located on the average face for each sequence and propagated to the individual frames. As shown in Figure 1, facial geometry can be represented by a Delaunay triangulation mesh. Texture information is sampled and warped to an average face shape by a piece-wise affine transformation, providing 5000 samples (pixels) from each frame.

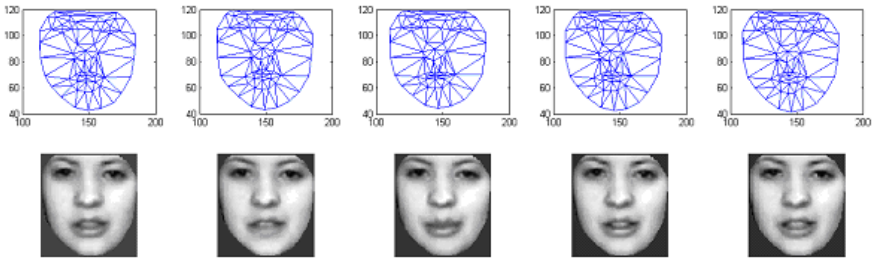


Fig. 1. After tracking facial movements, the face instance in each frame is triangulated and warped to an average shape reference for modeling

3.2 Facial Modeling

A combined shape and texture model [15] provides a linear statistical simulation of the human system. This has become popular for psychological facial analysis applications (e.g. [16]) and models both geometric and texture variations. If the feature points $X_h = (x_h, y_h)$ are expressed as a vector $\mathbf{S}_i = (X_1 ... X_h ... X_N)$, the shape can be represented as a combination of feature weights,

$$\mathbf{s}_i = {\mathbf{\Phi_S}}^T(\mathbf{S}_i - \bar{\mathbf{S}}) \tag{1}$$

where the $\mathbf{\Phi_S}$ eigenvectors correspond to the m largest eigenvalues derived from a Principal Components Analysis (PCA) of a suitable ensemble, and $\bar{\mathbf{S}}$ is the mean face shape. The texture features $\mathbf{t}_i$ are obtained by sampling intensity values $\mathbf{T}_i$ across the mesh triangles, followed by coding,

$$\mathbf{t}_i = {\mathbf{\Phi_T}}^T(\mathbf{T}_i - \bar{\mathbf{T}}). \tag{2}$$

It is still possible however, that there are redundancies between geometric and texture features, and these are reduced by another encoding,

$$\mathbf{c}_i = {\mathbf{\Phi_C}}^T \left(\begin{bmatrix} \mathbf{t}_i \\ w\mathbf{s}_i \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{t}} \\ w\bar{\mathbf{s}} \end{bmatrix} \right) \tag{3}$$

where $\mathbf{\Phi_C}$ represents the eigenvectors of samples formed by concatenating the shape and texture features and w is a weighting factor, calculated using the total variance of the shape and texture, so as to equalize the importance of the two components.

3.3 Human Recognition Simulation

Our previous work used facial motion information to improve recognition [17]. When faces are encoded by the statistical model, the orientations of the most significant variations are used to provide dynamic signatures. However, this motion information is not usable if static images are shown in the learning or recognition stages because it cannot be matched between gallery and probe.

Either a sequence or a static image can be projected onto the subspace for assessing similarity with the gallery faces. Following psychological results [8], we use normalized correlation as a distance measure,

$$S_{ij} = \frac{\mathbf{f}_i}{|\mathbf{f}_i|} \cdot \frac{\mathbf{f}_j}{|\mathbf{f}_j|}, \tag{4}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ represent two feature vectors selected by distinctiveness level at learning of the gallery face. Dynamic faces are represented by taking the mean of the sequence parameters $\mathbf{c}_i$, before calculation of distinctiveness. S_{ij} is the cosine of the angle between vectors from the mean face to the probe and gallery faces, measured only on the high-variance gallery parameters. Note that S_{ij} and S_{ji} are likely to be calculated on a different number and selection of parameters.

The experiment is run in the same manner, using the same faces, as the psychological experiments. At the learning stage, 20 faces, either Caucasian or Japanese, presented either as single static images or as moving sequences (each 50 frames long), are encoded by the model and the parameters are recorded. At the recognition stage 20 familiar and 20 unfamiliar faces are encoded and the minimum S_{ij} found for each. ROC curves are calculated, and the Area under the Curve (AUC) found, providing a criterion-free measure of discrimination. The psychological experiment used a between-subject design, and so each training or testing group consists of a single condition.

Distinctiveness of the parameters is ensured by setting a λ-value criterion above which the parameter on each dimension must lie to be included in the training feature vector. The same mask is then used to select features during recognition. In addition, structural enhancement improvement of the localization of facial features is simulated by adding Gaussian noise to the feature-point locations of the static images. This is parameterized by the standard deviation of the noise, measured in pixels. This study uses 500 Caucasian face images randomly selected from the FG-NET database and 60 similar randomly selected Japanese face images to simulate the recognition system. Ten such models are built, allowing standard errors to be calculated.

3.4 Results

The effects of altering the distinctiveness cut-off are shown in Figures 2 and 3. Clearly, the value of the criterion has a major effect on the ordering of the AUCs, with a shift from a symmetric preference for the same conditions when all of the parameters are included, to a situation where there is a preference for dynamic training, but not for testing. In addition, there is an other-race effect, with notably lower AUCs for the Japanese faces.

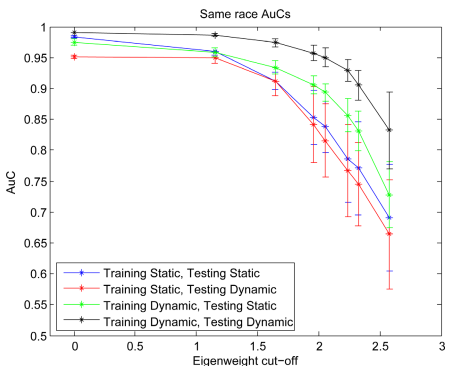


Fig. 2. Variation in AuCs for the Caucasian faces as a function of distinctiveness criterion

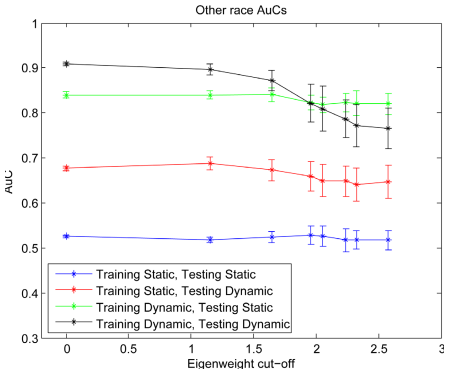


Fig. 3. Variation in AuCs for the Japanese faces as a function of distinctiveness criterion

These observations are backed up by Pearson correlations between the human AUCs and those from the model. These clearly show in Figure 6 that there is a larger variation in agreement for the Caucasian faces, and that peak correlation is found with a cut-off of approximately 2.24λ. The feature points found on the static faces were then distorted by different levels of noise, using a 2.24λ distinctiveness cut-off. The AUCs, given in Figures 4, 5 and, with Pearson correlations in Figure 7 show a decline in similarity of human performance for the Caucasian faces, but no such change for the Japanese faces, at moderate levels of noise.

4 Discussion

Considering the facial recognition results generated by this study, we find that this extended statistical model accords well with the human cognitive system. The findings can be summarized as follows: (1) motion effects do help improve facial recognition especially when dynamic information is shown in the learning stage; (2) better recognition for the same race faces provides further evidence for the existence of other race effects; (3) when facial parameters are selected at training by distinctiveness, this improves the relation with human performance

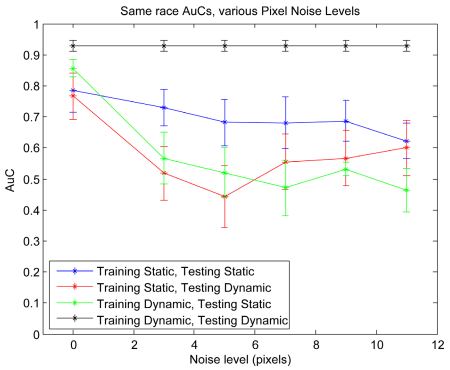


Fig. 4. Variation in AuCs for the Caucasian faces as a function of noise level

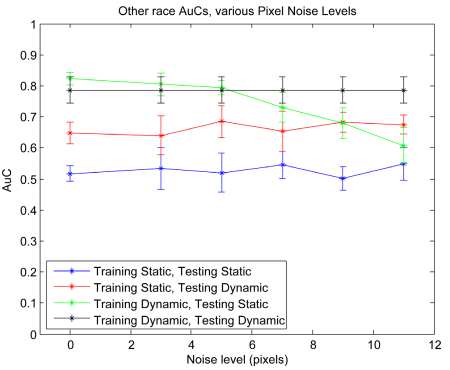


Fig. 5. Variation in AuCs for the Japanese faces as a function of noise level

for same-race faces, but not other-race faces; (4) decreasing the accuracy of the static-face interpretation reduces the realism of the same-race simulation, but not for the other-race faces.

These findings can be interpreted in terms of the representation enhancement hypothesis. The presence of additional views of the face allows more robust descriptions of faces. In addition, it appears that, when the face is of a type unfamiliar to the participant, more accurate description of the facial configuration becomes possible. Future work will consider processes involved in recognizing and tracking the individual features in human faces.

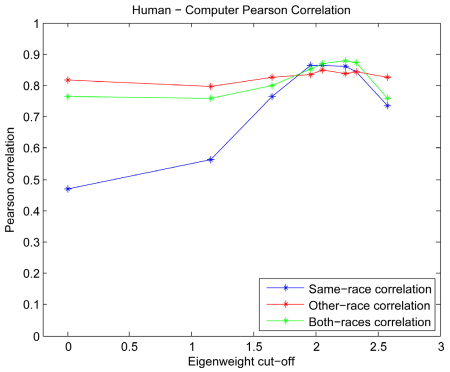


Fig. 6. Variation in correlation between Human and Algorithmic AUCs as a function of distinctiveness

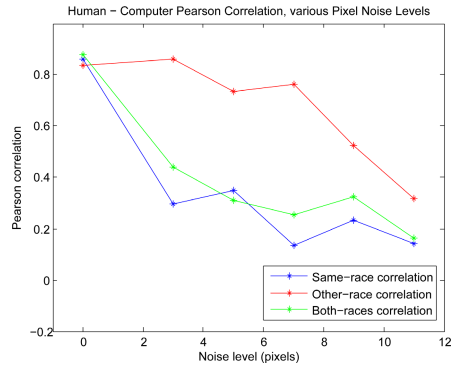


Fig. 7. Variation in correlation between Human and Algorithmic AUCs as a function of noise-level

References

[1] O'Toole, A., Roark, D., Abdi, H.: Recognizing moving faces: A psychological and neural synthesis. Trends in Cognitive Sciences 6(6), 261–266 (2002)
[2] Hancock, P.J.B., Bruce, V., Burton, A.M.: Recognition of unfamiliar faces. Trends in Cognitive Science 4, 330–337 (2000)
[3] Meissner, C.A., Brigham, J.C., Butz, D.A.: Memory for own- and other-faces: A dual-process approach. Applied Cognitive Psychology 19, 545–567 (2005)
[4] Tzou, C.H.J., Giovanoli, P., Ploner, M., Frey, M.: Are there ethnic differences of facial movmeents between europeans and asians? Surgical Reconstruction 58, 183–195 (2005)
[5] McKone, E., Brewer, J.L., MacPherson, S., Rhodes, G., Hayward, W.G.: Familiar other-race faces show normal holistic processing and are robust to perceptual stress. Perception 36, 224–248 (2007)
[6] Lander, K., Bruce, V.: The role of motion in learning new faces. Visual Cognition 10, 897–912 (2003)
[7] Dakin, S.C., Omigie, D.: Psychophysical evidence for a non-linear representation of facial identity. Vision Research 49(18), 2285–2296 (2009)
[8] Hill, H., Claes, P.D.H., Corcoran, M., Walters, M., Johnston, M., Clement, J.G.: How different is different? Criterion and sensitivity in face-space. Frontiers in Psychology 2 (2011)
[9] Valentine, T.: Face-space models of face recognition. In: Wenger, M.J., Townsend, J.T. (eds.) Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges, pp. 83–113. LEA, Mahwah (2001)
[10] Lewis, M.B.: Face-space-R: Towards a unified account of face recognition. Visual Cognition 11(1), 29–69 (2004)
[11] Butcher, N., Lander, K., Fang, H., Costen, N.: The relative effect of motion at encoding and retrieval for same and other race face recognition. British Journal of Psychology 102(4), 931–942 (2011)
[12] Marr, D.: Vision: A computational investigation into the human representation and processing of visual information. W. H. Freeman, San Francisco (1982)
[13] Rhodes, G.: Looking at faces: first-order and second-order features as determinants of facial appearance. Perception 17, 43–63 (1988)
[14] Wang, F., Vemuri, B., Rangarajan, A., Schmalfuss, I., Eisenschenk, S.: Simultaneous nonrigid registration of multiple point sets and atlas construction. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(11), 2011–2022 (2008)
[15] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transations on Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)
[16] Ashraf, A., Lucey, S., Cohn, J., Chen, T., Ambadar, Z., Prkachin, K., Solomon, P.: The painful face - pain expression recognition using active appearance models. Image and Vision Computing 27(12), 1788–1796 (2009)
[17] Fang, H., Costen, N.P.: From rank-N to rank-1 face recognition based on motion similarity. In: Cavallaro, A., Prince, S. (eds.) British Machine Vision Conference (2009)

An Audiovisual Feedback System for Pronunciation Tutoring – Mandarin Chinese Learners of German

Hongwei Ding[1], Oliver Jokisch[2], and Rüdiger Hoffmann[2]

[1] School of Foreign Languages, Tongji University, Shanghai, China
hongwei.ding@tongji.edu.cn
[2] Institute for Acoustics and Speech Communication, TU Dresden, Germany
{oliver.jokisch,ruediger.hoffmann}@tu-dresden.de

Abstract. Computer-assisted pronunciation tutoring (CAPT) methods have been established during the last decade. Recent systems usually include a distinct user feedback and an automatic pronunciation assessment system. This study is based on the audiovisual CAPT system, in which an extensive feedback mechanism and several speech databases for Slavonic learners of German were developed. We intend to adapt the existing system for Chinese learners of German and report on the first usage experiences. We have thus analyzed the deviations of German utterances produced by Chinese learners in comparison to those of German natives, especially in term of prosodic and phonetic issues. We also designed supplementary database and organized perceptual evaluation tests by German native listeners with respect to individual phones as well as to general rhythm and intonation. In this way the language transfer of tonal Chinese can be demonstrated, which is vital to the system adaption for Chinese learners.

Keywords: Computer-assisted pronunciation tutoring, learning German, Mandarin Chinese learners.

1 Introduction

With the progress of speech technology, language educators become more interested in Computer Assisted Pronunciation Training (CAPT). Many complete pronunciation tutoring systems have been developed for foreign language learning, among which there are systems for Cantonese learners of English [1], for Japanese learners of English [2] and for German learners of Chinese [3]. All these systems intend to capture segmental and suprasegmental error patterns of the learners and provide the learners with targeted training to improve their pronunciation. Following the same approach, we will focus on a German learning tutoring system EURONOUNCE [4] in this paper, which is an Intelligent Language Tutoring System with multimodal feedback functions, a project funded with the support from European Commission. EURONOUNCE is a corpus-based learning system, which integrates large speech corpora and multilingual speech databases. Special speech corpora are needed for each pair of languages. Language pairs include German/Russian, German/Polish, German/Czech, etc. With the aim to extend the system to German/Chinese language pair, we have been making efforts to collect speech data and analyze the learning effects of Chinese students. This paper will report the first experience to build such a tutoring system.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 191–197, 2012.

In the following two sections, our efforts to adapt the system for Chinese learners will be illustrated, and the benefits it brings to Chinese learners and problems for an effective use of the tutoring system will be summarized.

2 System Extension for Chinese Learners of German

In order to adapt the system for Chinese students we will first describe the characteristics of Chinese learners, then illustrate our work plan, and finally demonstrate the phonetic and prosodic difficulties of Chinese students.

2.1 Characteristics of Chinese Learners

With an analysis of Chinese learners of German, there are three main arguments for the application of CAPT in China:

1) Foreign language classes in China tend to emphasize reading and writing more than speaking, pronunciation in L2 learning is generally neglected. One reason is that the low teacher-student ratio does not allow individual pronunciation exercises. Another problem is that teachers are embarrassed because of the lack of phonetic instruction strategies. Computers can thus offer a solution to this problem, by engaging the students in one-to-one pronunciation exercises and providing feedback and suggestions for improvements.
2) As a lingua franca, English is taught as the first foreign language in schools in China. Most students begin to learn German as a second foreign language at universities when they are over 18 years old. Their perceptual discrimination of phonetic sounds is not as good as that of a child; their learning of pronunciation should be enhanced by informative visual feedbacks.
3) Chinese has a logographic orthography; phonological sensitivity is not associated with reading. Students with non-alphabetic L1 background will rely more on visual processing than on phonological processing in reading. The pronunciation deviations can, however, hardly be perceived by Chinese learners on their own. Tutoring systems with feedback information best fit the requirements of Chinese German learners.

The language educators in Chinese German College (CDHK: Chinesisch-Deutsches Hochschulkolleg) at Tongji University find CAPT systems especially helpful for their students. In CDHK postgraduate students without any knowledge of German language start with intensive German language training, and after two years' courses they will go to Germany for exchange program. But their heavy foreign accents present a great obstacle for the oral communication in Germany, because they can hardly distinguish different German sounds perceptually and productively. However it has been proved that with some degree of awareness and visual aids, the acquisition of a nearly accent-free pronunciation of German is still possible for these adult students.

As phonetic researchers, we share common interest with German language teachers to improve the pronunciation of Chinese students. We installed the EURONOUNCE tutoring software in the language lab of CDHK. On one hand, the students could have

access to the tutoring system to find their pronunciation errors and train their listening and pronunciation abilities; on the other hand, during their exercise the speech materials would be collected automatically in the computer. We work hand in hand with the language teachers to ascertain the difficulties and assess the progress of the students, and tried to compile a suitable curriculum and optimize the system especially for Chinese students.

2.2 Collection of Speech Data

Based on the baseline system and database infrastructure of the accomplished EURONOUNCE tutoring system, we still have to 1) collect speech data; 2) analyze the phonetic and prosodic deviations of the students.

Main source of the data was collected when the students used the tutoring system in their lab. The user interface of this system can be illustrated in Fig. 1, in which different colors show different degrees of accuracy assessed by the system automatically. In this example, phonemes with green colors are evaluated as acceptable, but /U/ in *Stubben* ("*stumps*") with orange color indicates that this sound should be improved. It is also demonstrated in our previous study [5] that Chinese students have some difficulty in producing short vowels such as /U/.

Fig. 1. User interface of the pronunciation tutoring system

Armed with the audiovisual feedback, students could compare their own sounds with those of the standard speaker, find out the discrepancies, and try their best to imitate the standard one. The speech data in the drill process of about 100 beginners have thus been collected for our research purpose.

In order to assess the effects of using the tutoring system, we designed questionnaires to collect information regarding phonetic pronunciation training before, during and after the use of the tutoring system, questions such as:

1) How do you rate your ability to speak German?
2) How do you rate your ability to understand native speakers in conversation?
3) Do you think it is important to know the sounds of German?
4) Do you use phonetics to help you understand how to say new vocabulary?
5) Can you hear the difference between the long and short vowels in German?

Such kind of questions have been asked at different stages in using the system.

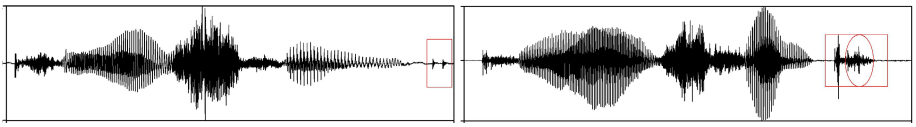
In order to supplement the exercises in the system, we also designed some database and made extra recordings. These speech utterances were further rated perceptually by German native listeners, so that we could find the correlation of acoustic discrepancies and perceptual foreign accents. The special difficulties of Chinese students could thus be accurately ascertained and acoustically demonstrated. Some results of our investigations have been reported [6], [7], [8].

2.3 Analysis of Phonetic and Prosodic Deviation

In this way, we have collected a large database, each student used the tutoring system 1 hour/week averagely. With about 100 students, we could collect about 100 hours of recording per week. But most of the utterances were only repetitions, which proves less worthy for phonetic research, but could shed some light on the progress and learning motivations. With acoustic and perceptual investigations, we have found that various segmental and intonational deviations accumulated to account for difficulties in oral communication. Some of them are listed:

1) Inaccurate production of those German vowels and consonants which are nonexistent in Chinese
2) Incorrect placement of tonal categories and wrong phonetic realization of a phonological category

Chinese learners usually employ different strategies, such as epenthesis, deletion and modification to deal with unfamiliar sounds. Because Chinese syllables usually end with vowels, and the learners usually add a schwa /@/ after the consonant final. In Fig. 2 a Chinese student added an /@/ (marked in ellipsis in (b)) after /k/, so that a native speaker's /k/(marked in rectangle) shown in (a) was replaced by a Chinese speaker's /k@/ (marked in rectangle) shown in (b). In a comparative analysis of German produced by Russian and Chinese learners, it is obvious that epenthesis occurs more frequently among Chinese learners than Russian learners [9].

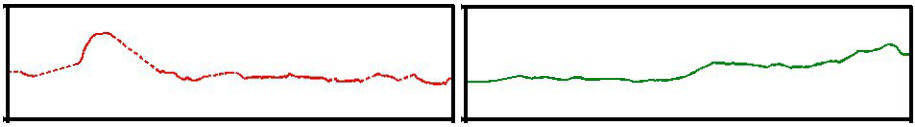


(a) (b)

Fig. 2. The waveforms of *Kühlschrank* (*icebox*) produced (a) by a native speaker; (b) by a Chinese learner

With the visual-audio feedback information and after many times of trial and error, the learners became conscious of their pronunciation mistakes, and could make correspondent corrections.

Chinese is a tone language, Chinese speakers thus raise or lower their pitches to express different lexical meanings instead of different linguistic purposes in intonation languages like German and English. Previous findings in f0 deviations of Chinese speakers of German have been investigated in [10]. Another example in the current study is shown in Fig. 3, where a Chinese student could not raise their pitch at the end of the sentence to indicate a question shown in (a). But with the guidance of f0 contours, the student could finally manage to imitate the question intonation described in (b).



(a) (b)

Fig. 3. F0 contours of the question "*Hast du etwa Bohnen in den Ohren?* (*Have you any beans in your ears?)*". (a) F0 produced by a student at the beginning of exercise (b) F0 produced by the same student after some exercises with the tutoring system.

Visualization of intonation curve is proved to be particularly effective in the acquisition of L2 intonation. Automatic pitch tracking algorithm, however, usually displays many small pitch changes that make the learners confused about the sentence intonation, and moreover these small changes are linguistically unimportant. Some tools should thus be applied to smooth the intonation curves, so that only perceptually relevant pitch changes are displayed, which can greatly facilitate the learners' acquisition of intonation.

2.4 Advantages and Problems with the Tutoring System

It is obvious that with the help of audiovisual feedback information, students can gradually acquire the right phonological categorization in perception and master the sensorimotor skills for production. But there are still some practical obstacles that should be overcome to facilitate this process, some of them are listed below:

1) It seems that the technique to generate the learner's voice with the native speaker's pitch contour for feedback can best facilitate the acquisition, which should be integrated into our system.
2) If the students share the same language lab when using the tutoring system, a quiet environment can hardly be guaranteed, which will influence the accuracy of speech recognition. The feedback information is then not reliable due to wrong assessment results. If they use the tutoring system at home, the data can hardly be collected.
3) Most of the language teachers have little knowledge of acoustic phonetics, they find it difficult to understand the information provided by the system. They can hardly interpret the correlation between the perceived wrong pronunciation of the input

and the visual acoustic output, an introductory phonetics course is indispensable for the language educators.

4) The students wish that the reading material stored in the system can change flexibly, ideally the material can accompany their textbooks. But the recognition system seems not powerful enough to provide an accurate assessment of individual phones in long and continuous speeches.

In order to optimize the system and best fit the requirements of the learners, phonetics researchers, speech technology experts and language educators should deal with these problems together.

3 Conclusion

It proves possible for Chinese learners of German language to imitate standard pronunciations successfully. However, a faithful imitation of isolated words or sentences with visual aids can not guarantee a good pronunciation in ordinary speech. The articulatory constraints will still dominate for many students in normal speech without any audiovisual aids. The tutoring system should also guide the learners step by step from a successful imitation to an accurate production in free continuous speech. Therefore the next research interest will be focused on continuous and normal speech, which is the ultimate goal of language teaching.

Acknowledgments: The first author is sponsored by Shanghai Pujiang Program (Project No. 11PJC099) and Shanghai Social Science project (Project No. 2011BYY002) for this research work.

References

1. Qian, X., Meng, H., Soong, F.: Capturing L2 Segmental Mispronunciations with Joint-sequence Models in Computer-Aided Pronunciation Training (CAPT). In: Proc. Chinese Spoken Language Processing (ISCSLP), pp. 84–88 (2010)
2. Minematsu, N., Okabe, K., Ogaki, K., Hirose, K.: Measurement of Objective Intelligibility of Japanese Accented English Using ERJ (English Read by Japanese) Database. In: INTERSPEECH 2011, pp. 1481–1484 (2011)
3. Hussein, H., Mixdorff, H., et al.: Towards a Computer-aided Pronunciation Training System for German Learners of Mandarin - Prosodic Analysis. In: L2WS-2010, Tokyo (2010)
4. Jokisch, O., et al.: The EURONOUNCE Project - An Intelligent Language Tutoring System with Multimodal Feedback Functions: Roadmap and Specification. In: Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Frankfurt, pp. 116–123 (2008)
5. Ding, H., Jokisch, O., Hoffmann, R.: Perception and Analysis of Chinese-accented German Vowels. Archives of Acoustics 32(1), 89–100 (2007)
6. Ding, H., Mixdorff, H., Jokisch, O.: Pronunciation of German Syllable Codas of Mandarin Chinese Speakers. In: Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Berlin, pp. 281–287 (2010)
7. Ding, H., Jokisch, O., Hoffmann, R.: An Acoustic and Perceptive Analysis of Postvocalic/l/ in Mandarin Chinese Learners of German. In: 17th Intern. Congress of Phonetic Sciences (ICPhS), Hongkong (2011)

8. Jokisch, O., Ding, H.: Acoustic Analysis of Postvocalic /l/ in Chinese Learners of German in the Context of an Overall Perception Experiment. In: Proc. Speech and Language Technology in Education (SLaTE), Venice (2011)
9. Hilbert, A., Mixdorff, H., Ding, H., Pfitzinger, H., Jokisch, O.: Prosodic analysis of German produced by Russian and Chinese learners. In: Proc. 5th Intern. Conf. on Speech Prosody, Chicago (2010)
10. Ding, H., Jokisch, O., Hoffmann, R.: F0 Analysis of Chinese Accented German Speech. In: Proc. 5th Intern. Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, pp. 49–56 (2006)

Si.Co.D.: A Computer Manual for Coding Questions

Augusto Gnisci, Enza Graziano, and Angiola Di Conza

Department of Psychology, Second University of Naples,
Viale Ellittico, 31 - 81100 Caserta, Italy
{augusto.gnisci,enza.graziano,angiola.diconza}@unina2.it

Abstract. This contribution aims at presenting a computer manual for coding questions called Si.Co.D. (standing for "Sistema di Codifica delle Domande", that is Coding System of Questions). The software presents a set of related coding systems that identify questions on the basis of their openness/closeness, threatening, confusing formulation and intonation. The software has the characteristics of interactivity and multimediality, in order to facilitate and help observers' training for coding questions. It is composed of two sections: one dealing with flow chart and definitions of questions categories, the other one presenting some examples of questions. Peculiarity and advantages of the tool are described, also comparing it with other annotation and sound management software. Possible applications are discussed. Its usefulness is showed by some research applications.

Keywords: Si.Co.D. software, Coding systems, Questions, Training.

1 Introduction

The present contribution aims to propose a free and easy software conceived to support observers' training in learning new coding systems about questions. Indeed it favours the understanding of basic constructs to be applied to the analysis of questions and of the operationalization on which each construct and each category is based. The software is called Si.Co.D.[1] (standing for Sistema di Codifica delle Domande, that is System for Coding Questions). It consists of a computer manual for coding questions according to their openness/closeness form (which is connected with their coerciveness), threatening formulation, confusing construction and intonation. Therefore, it deals with the analysis of interaction and, in particular, it has been conceived to be applied to interactions based on the question-answer exchange, a format occurring frequently in different contexts, such as courtroom, classroom, interviews (including political interviews). Analysis of interaction and conversation is important for the understanding of cognitive behavioural systems and their implementation. Therefore, creating and sharing systems and tools which facilitate the coding of some features of interaction and allow its analysis and knowledge, is fundamental. Si.Co.D. constitutes one of this tools, because it helps researches dealing with a formal feature of conversation, that is the question-answer exchange.

[1] Si.Co.D. is a freeware downloadable at http://osservazione.co.cc/Download_Software.html

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 198–207, 2012.

The recourse to a categorization process is known to be a relevant human cognitive process; furthermore the use of formalized category systems is frequent in research. For example, in observational research, categories and category systems represent important tools, through which reality can be known and structured information can be collected [1]. In psychological research, behavioural categories are often applied to the study of human behaviour, and studies testing the identified categories are generally conducted in order to constitute the corresponding category system, usually checked for usability and reliability [1-2].

Once a category system has been structured and tested, it becomes a useful tool for conducting other studies on similar matters; then observers need to be trained in order to properly associate categories to events. Thus they have to understand and learn conceptual definitions of each category and to identify the right association between them and the reality. This process can be substantially supported by referring to the whole category system and by the recourse to real examples of each category (via computer and via software).

Dealing with observers' training and inter-observer reliability, Si.Co.D. also represents a support to observational research [1-2]. Indeed it facilitates coders' training on different sets of coding systems concerning different questioning characteristics.

In the next paragraphs, we will explain what a question is, how it can be characterized and we will describe the coding systems of questions presented by Si.Co.D.. Moreover, the developmental process of the software and its interface and characteristics will be highlighted.

2 The Coding System of Questions

According to a functional perspective, a question is defined as a request for information [3-4]. The request can be formulated in several different ways. Si.Co.D. stores four coding systems of questions, derived from the literature on the theme [5-10] that are linked to the openness/closeness of a question (which deals with its coercion), face threatening, confusion and intonation. For each category system, both micro and macro-categories adapted to Italian language are presented. The micro-categories can be identified in different languages considering their specific grammatical construction of a question. Because of the relation existing between linguistic characteristics and micro-categories, the comparison between different languages can be done only considering the macro-levels of coding system, that are less linked to the specific language and more associated to psychological concepts. The category systems are described in the next paragraph.

2.1 Description of the Coding System

Four coding systems developed in Italy and in other countries (then adapted to Italian) [5-10] are included in Si.Co.D.. openness/closeness, face threaten, confusing formulation, intonation of questions.

Open/closed questions. The openness of a question concerns the degree of freedom leaved to the respondent when providing the requested information. Closed question reduces the choice and the length of response, implicitly carrying a set of presuppositions and inducing the desired response [6, 8]. The openness of a question is strictly related to its coerciveness, that is the way in which the form of the question constrains the answers [8-9], 11,12] and imposes the interviewer's version of the events [11]. For this system we can identify two macro-levels: the *closed questions*, including the micro-categories labelled statements, yes/no questions and tag questions; and the *open questions*, including the micro-categories named wh-questions, introduced by interrogative pronouns. As briefly mentioned above, the recourse to the macro distinction linked to the openness/closeness categorization allows the cross-cultural comparison of questions formulated in different languages (see below). The micro-categories are based on grammatical and intonation characteristics of the question and can be applied to questions formulated in every kind of language. Thus, the recourse to the micro-categories does not permit the comparison of the frequency of use between different languages. On the contrary, this comparison is allowed by the macro-level of openness/closeness of questions (see par. 4).

Threatening questions. A question is considered threatening when each and every possible answer carries a damage to the respondent's "face" [13] as a potential consequence. The "face" is conceptualized as the desirable image everyone wants to give of him/herself to the others in terms of social attributes [14]. The face can be lost, preserved or improved [15]; some behaviours, acted out from one's interactant, can represent a threat to the face, for example expressing disagreement [16]. This kind of questions is often used in those contexts where providing a positive image of him/herself is paramount. For example in political interviews, the interviewer often threatens the face of the politician and the politician must be able to avoid the threat to preserve, or even improve, his/her own face [13, 16-18]. For identifying whether a question is or is not threatening, the basic form of the question must be identified and then all the possible answers must be picked out [7], in order to establish if there is at least one answer that, once provided, does not represent a risk to lose one's own face. If it exists, the question must be coded as not necessarily threatening the face of the respondent; whether, on the contrary, all the possible answers are damaging, the question must be coded as threatening [13]. Studying the political context, 19 types of threatening questions have been identified [13], belonging to three superordinate categories, according to which feature of the respondent is threatened: his/her face, the face of the his/her party and the face of his/her significant others (e.g., in the case of a politician, colleagues or allied).

Confusing questions. Questions expressed in a complex and unclear way are defined "confusing". The confusing formulation of a question derives from lexical or grammatical characteristics. Questions which are negative, double negative, leading, multiple, with complex syntax, with complex lexicon are defined confusing [10].

Intonation. Finally, the questions intonation is the last feature reported by Si.Co.D. Intonation concerns the tone change in conclusion of a question. So we can have questions with a falling, constant or rising intonation [19].

3 The Development of the Software

In observational research, the correspondence between the categories and the real events must be learnt and understood in order to produce a correct coding of the observed data and to gain founded conclusions. That's the reason why coders are generally trained to use and apply a category system. In this phase observers may often have difficulties in learning conceptual definitions of each category or in discriminating between categories. Si.Co.D. comes out primarily of the awareness of these problems. Indeed, the use of a software like Si.Co.D. can favour a better outcome and a greater homogeneousness of the training.

Si.Co.D. is composed of two sections: in the first one each category is defined and described; in the second one some examples are provided for each type of question.

A business developmental software, called Macromedia Flash 8 Professional was chosen to program the software in order to get two fundamental characteristics: interactivity and multimediality.

Interactivity allows to deepen information on the categories in a dynamic way: coders can easily read the description of each category and see and listen to the audio- or video-recorded examples of the questions. Thanks to this support, the coding system can be learnt quickly and correctly because users can decide what to do in a specific moment: they can move from one section to another one according to their own demands. For instance, they can choose to see a flow chart of a macro-level of categories, or have a definition of a certain type of question, or listen to an example.

Si.Co.D. is characterized as a multimedia instrument as well. It means that the software is able to show and manage multimedia flows, including audio and video materials (namely, the examples of questions). Thanks to this feature the learner can go through a better simulation of the observation, especially for what concerns the coding phase.

Interactivity and multimediality make Si.Co.D. an easy-to-use software.

3.1 How Si.Co.D. Is Made

As above-mentioned, Si.Co.D. is composed of two sections: the first one describes the flow chart for coding questions and the other one provides a catalogue of examples. Users can reach each section simply by a click of the mouse.

When the software is working, the user can see a short description of Si.Co.D. and then can choose what to do through the menu, placed on the left side of the screen. The software presents four category systems, corresponding to the macro-levels (openness/closeness, threatening, confusing construction and intonation of questions) and for each category system its categories are listed and described (micro-levels based on grammatical and intonation characteristics of questions). Si.Co.D. consents to visualize both the levels, which can be reached with a mouse click. Each page shows a macro-level with the corresponding flow chart and the respective categories. When the mouse overlaps the denomination of a category, a popup appears in the left part of the screen, containing the definition of the category and the main information about it. Providing these essential characteristics of the category allows to identify and understand it correctly, leading to a consequent proper use.

The second section of the tool is the catalogue of examples of each question, and is organized in macro-and micro-levels of questions as well, following the organization of the whole coding system. Clicking on the macro-levels (openness/closeness, threatening, confusing construction and intonation of a question), lower level categories appear in an upper central window. Similarly, clicking on one of the lower level categories, other lower categories could potentially appear on the right side of the screen. At the same time, in the middle window all the available examples will appear. The user can click on one of the audio- or video-recording of a question and watch or listen to it. While the example is going on, he/she can contemporarily read the transcription of the text which appears on the screen.

In sum coders can learn conceptual definitions of categories and have examples of them in the meanwhile.

Coming from different observational and interactive contexts, the examples have been accurately sampled among more than twelve hours of audio- or video-recorded question-answer exchanges during courtroom interactions, political interviews, panel discussions, school interactions. Details are provided in Table 1.

Table 1. Examples of questions reported in Si.Co.D. as in [23] Remark 1. * Creative Commons, "Attribuzione 2.5 Italia".

Name	Duration	Date
"Processo La Mantia ed altri" (radio radicale.it*)	1.33.46	9th June, 2008
"L'Ue e gli Usa nella crisi: intervista ad Antonio Panzeri" (radio radicale.it*)	0.06.41	11th March, 2009
"Processo a carico di Giovanni Mercadante ed altri" (radio radicale.it*)	1.19.52	23th April, 2009
"Processo Occidente (Altadonna ed altri)" (radio radicale.it*)	0.23.30	1th October, 2008
"Processo Occidente (Altadonna ed altri)" (radio radicale.it*)	2.47.26	16th October, 2008
"Processo Grauso ed altri" (radio radicale.it*)	4.02.04	14th October, 2008
"Intervista a M. Capano su Registro Testamenti Biologici" (radio radicale.it*)	0.14.31	9th March, 2009
"Omicidio Vincenzo Casillo" (radio radicale.it*)	1.30.00	19th December, 1988
"In Mezz'Ora" (Rai 3)	0.32.00	29th June, 2008
"L'Era Glaciale" (Rai 3)	1.39.00	24th April, 2009
"In Mezz'Ora" (Rai 3)	0.29.48	31th May, 2009
Total duration	12.46.46	

To make the software more user-friendly, every page is associated with an on-line guide, available through the symbol "?". A popup will explain how to use Si.Co.D.

4 Using Si.Co.D.: Peculiarity and Advantages

Effective research tools, based on adequate and reliable coding systems can offer a standard shared by the researchers operating in the same field. Researchers studying the same phenomenon often use different "languages" and instruments, that make their studies not comparable. On the contrary different studies and diverse hypotheses can be compared creating an exhaustive system, useful for many intents and usable by many researchers who can finally cooperate and interact. In this sense, Si.Co.D. was conceived for providing a standard for coding questions.

In our opinion, ecological examples of the categories, reproducing questions really posed by an interviewer, represent one of the main advantages of this software. The examples are fruitful because they allow the coders to practise a real coding process. Moreover they show directly how the constructs underlying each category system has been operationalized, improving the understanding of the correspondence between categories and events and of the psychological meaning of the constructs.

Another advantage of Si.Co.D. is the opportunity to conduct cross-cultural comparison by the use of macro-categories (such as the openness/closeness of questions, related to their coercion). The next example illustrates this aspect. In English, a question is generally introduced by an auxiliary verb; or, any case, its formulation follows some fixed rules, like the inversion of the subject with the verb. In Italian, the grammatical structure of an affirmative sentence and of a question can be the same, but the intonation varies. As a consequence, the frequency of use of a certain type of question (for example, of yes/no questions, which constitute a micro-level) between English and Italian interviewers can not be compared. Indeed, at this level the differences among the two interviewers are not due to their characteristics, but to differences linked to their own mother-tongue. However we can compare the two interviewers by considering the frequency of use of the super-ordinate categories (macro-categories) defined at a broader level and including the micro-levels. Going back to the example, in the specific case of yes/no questions, we can compare the openness/closeness dimension, that, as previously described, is linked to the concept of coercion and rests on the clustering of the micro-categories. Both in Italian and in English (as well as in several other languages), the dimension of openness/closeness of the questions can be identified, allowing to compare different languages. This is possible because this distinction is not based on grammatical differences but on a broader psychological meaning, linked to the construct of coercion. So, if English interviewers use more frequently one over the other micro-categories than Italian interviewers, we cannot gain any comparative conclusion. On the contrary, if English interviewers use more closed questions than Italian ones, then we are allowed to conclude that the first ones exert more control on the conversation than the latter ones.

Moreover, noteworthy Si.Co.D. is not a closed software; on the contrary, it is flexible and adaptable, it can be modified in order to add new categories with relative examples, and suggestions coming from future research, particularly from different languages, can be enclosed.

Finally, being Si.Co.D. easy to use, there is no need for a specific training on its usage. The online-guide appearing at each page helps users to solve potentially arousing doubts.

4.1 Differences between Si.Co.D. and Other Tools

Si.Co.D. has been implemented to be a computer manual for coding questions. It deals with a coding system based on four coding systems of questions linked one to another (see above). Using Si.Co.D. during coding means having at one's disposal: a flow chart representing macro- and micro-levels of categories; conceptual definitions of each category; the differences among them; real examples of questions taken from "natural" settings which operationalize the constructs; the written transcription of each example.

Because of its main aims, Si.Co.D. differs from other popular tools, like ELAN (www.lat-mpi.eu/tools/elan/) and ANVIL (my.tbaytel.net/tgallo/anvil/), in topics, functions and aims. The last ones are annotation tools whose function is to allow the annotation of many types of audio- and video-streams. ELAN, for instance, allows the flexible management of annotation files, for example by associating an annotation file to more than one audio- or video-records and annotations can be created on multiple layers. Si.Co.D. is a manual, not an annotation tool: it provides audio- and video-records, but they represent examples of categories of questions supporting the coders' learning and training phase.

Another spread annotation tool is ANVIL. Using this software, observer can code in real time while the video goes on, having at his/her disposal some functions helping the coding process. So it can be applied in many fields. On the contrary, Si.Co.D., although dealing with questions coding, does not want to manage audios or videos to be coded in real time. As mentioned above, it proves useful before the real coding process, that is when studying and learning the coding systems of questions. Being a computer manual, it provides definitions of categories and coding examples, but it does not intend to favour the computerized coding process of observational data.

There are many used software which deal with sounds and connected operations, such as WAVESURFER (sourceforge.net/projects/wavesurfer/) and PRAAT (www.fon.hum.uva.nl/praat/). The former deals with analysis and manipulation of sounds. The basic audio editing operations of modification of sounds and transcriptions are allowed. So, applied to our research field, the joint usage of WAVESURFER and Si.Co.D. could prove useful, in order to analyze pitch or intonation of questions or to transcribe the text of an observed interaction. However, it is not a specific tool dealing with questions and it doesn't provide any coding system such as Si.Co.D. does. The same considerations can be done for PRAAT, that is another tool concerning phonetics. It deals with analysis and reconstruction of acoustic speech signals and it has functions such as leaning algorithms and statistics. On the contrary Si.Co.D. has a more basic and specific function, dealing with a formal feature of speech, that is the identification of questions characteristics in some interactional contexts.

5 Researches Supported by Si.Co.D.

Di Conza et al. [16] tested the efficacy of Si.Co.D. in a study dealing with the interviewing style of two Italian journalists towards two opposite political parties. In particular, this software was used to code coercive questions asked to the representatives of the two political parties before and after two elections (2004 European and 2005 Regional elections). They coded 102 broadcasts providing 1391 question-answer sequences. In this study, Si.Co.D. was used for coding coerciveness and proved useful to favour coders' training and subsequent calibration, as showed by the "excellent" value of inter-observer reliability obtained in the study (Cohen's κ=.87). [20]

Before Si.Co.D. was implemented, some researches were conducted for coding questions by means of the coding systems described previously and presented in Si.Co.D.. Results of these studies were collected and analysed in order to conceive a

software dealing with a comprehensive coding system of questions, structured as a computer manual for coding questions. For instance, the category system concerning coerciveness was derived from Gnisci and Bonaiuto [21], who coded coerciveness of questions posed towards Italian politicians in political interviews and in courtroom, in a sample of 755 question-answer sequences during more than 4 hours of interviews (4 hours 31 minutes and 45 seconds). Cohen's *k*s were comprised among .66 and .89 and considered good/excellent according to Fleiss's [20] suggestions. Moreover, Gnisci [12] coded questions formulated by interviewers towards Italian left and right-wing politicians in a sample of 48 interviews lasting about 11 hours (11 hours 27 minutes). The level of coerciveness and face-threaten of the questions were analysed, showing a good inter-observer reliability (Cohen's k=.80 for coerciveness questions; k=.68 for threatening questions). These category systems were also employed to compare Italian and English journalists. In another study [22] dealing with the comparison between political interviews aired on different channels in 2006 and 2008 Italian General elections (the whole duration of the sample was 82 hours 12 minutes, identifying 2215 questions), threatening questions were coded obtaining an excellent [20] Cohen's *k*, both in the sample of 2006 (k=.93) and in the sample of 2008 (k=.75).

For what concerns Si.Co.D., the software has been used for training observers to code questions formulated in different languages (Italian and Greek) and in different contexts (courtroom and political interviews) in the course of on-going studies (not yet published).

In one of this studies, Si.Co.D. provided a useful support in coders' training for identifying questions used by interviewers in Italian and Greek political interviews aired during the election campaign of European election of 2009. 10 hours of Italian interviews and 10 hours of Greek ones have been coded, identifying 802 question-answer sequences. According to one of its fundamental aims, the software proved useful to compare questions formulated in different languages and cultural contexts and, consequently, to compare interviewers' features in conducting conversation. Inter-observer reliability was fair for threatening questions (Cohen's k=.58), and good for intonation (k=.63) and openness/closeness (k=.74). Moreover, this research provides an application for Si.Co.D. usage in a study dealing with cross-cultural comparison.

Another study, conducted with the support of Si.Co.D. aimed to compare questions formulated by interviewers towards Italian politicians in two different contexts. Questions addressed to the same Italian politicians were analysed both in courtroom and in political interviews. Overall, 37 hours of interaction were coded (37 hours 27 minutes), analysing 2758 questions (955 questions in political interviews and 1803 in courtroom). Inter-observer reliability was good (for openness/closeness, intonation and confusing formulation, Cohen's $.66<k<.79$) in this study, too. In addition, this research shows the application and the utility of the tool in two different contexts, where the interaction is characterized by the question-answer format.

6 Conclusions

Si.Co.D. provides practical advantages like consistency of training and standardization of practise for coders interested in analysing features of questions. The studies described in this contribution underline the utility of Si.Co.D. in coders'

training and calibration, as showed by inter-observer reliability. The tool is based on a set of coding systems of questions adapted to the Italian language, but usable in every language; moreover it allows to compare studies of different languages if the attention is posed to the macro-levels of categories based on psychological concepts, as showed by some researches cited in this article.

In conclusion, the proposed software has the following advantages: it consents a quick and accurate training of the researchers operating in the same or in similar fields and interested in studying the characteristics of the questions; it is easy to use, handy, intuitive and endowed of an user friendly interface; finally, it is flexible and projected to be enriched by new categories and examples, coming from other fields or from future studies on linguistic, communication and interaction.

References

1. Bakeman, R., Gottman, J.M.: Observing Interaction. An Introduction to Sequential Analysis, II edn. Cambridge University Press, New York (1997)
2. Bakeman, R., Gnisci, A.: Sequential Observational Methods. In: Eid, M., Dieneer, E. (eds.) Handbook of Multimethod Measurement in Psychology, pp. 451–470. American Psychological Association, Washington, DC (2005)
3. Bull, P.: On identifying questions, replies and non-replies in political interviews. J. Lang. Soc. Psychol. 13, 115–131 (1994)
4. Gnisci, A.: Le domande nella conversazione legale dialogica: Proposta di una tassonomia basata su criteri sintattici e intonazionali. R I L A 2, 45–80 (2000)
5. Adelswärd, V., Aronsson, K., Jönsson, L., Linell, P.: The Unequal Distribution of Interactional Space: Dominance and control in courtroom interaction. Text 7, 313–346 (1987)
6. Atkinson, J.M., Drew, P.: Order in court: The Organization of Verbal Interaction in Judicial Settings. Macmillan, London (1979)
7. Bull, P.: The Microanalysis of Political Communication: Claptrap and Ambiguity. Routledge, London (2003)
8. Matoesian, G.M.: Reproducing Rape: Domination through Talk in the Courtroom. University of Chicago Press, Chicago (1993)
9. Woodbury, H.: The strategic use of questions in court. Semiotica 48, 197–228 (1984)
10. Kebbell, M.R., Johnson, S.: Lawyers' Questioning: The Effect of Confusing Questions on Witness Confidence and Accuracy. Law Human Behav. 24, 629–641 (2000)
11. Danet, B., Hoffman, K.B., Kermish, N., Rafn, H.J., Stayman, D.G.: An ethnography of questioning in the courtroom. In: Shuy, R.W., Shnukal, A. (eds.) Language Use and the Uses of Language, pp. 222–234. Georgetown University Press, Washington, DC (1976)
12. Gnisci, A.: Coercive and Face-Threatening Questions to Left-wing and Right-wing Politicians During Two Italian Broadcasts: Conversational Indexes of Par Conditio for Democracy Systems. J. Appl. Soc. Psychol. 38, 1179–1210 (2008)
13. Elliott, J., Bull, P.: A Question of Threat: Face Threats in Questions Posed during Televised Political Interviews. J. Community Appl. Soc. 6, 49–72 (1996)
14. Goffman, E.: On Face-Work: An Analysis of Ritual Elements in Social Interaction. Psychiatry 18, 213–231 (1955); Reprinted in Goffman, E.: Interaction ritual: Essays on Face to Face Behavior, pp. 5–45. Anchor, Garden City NY (1967)
15. Brown, P., Levinson, S.C.: Universals in Language Usage: Politeness Phenomena. In: Goody, E. (ed.) Questions of Politeness, pp. 53–310. Cambridge University Press, Cambridge (1978)

16. Di Conza, A., Gnisci, A., Caputo, A.: Interviewers' Use of Coercive Questioning during a Midterm Period Favorable to the Opposition Party. In: Esposito, A., Esposito, A.M., Martone, R., Müller, V.C., Scarpetta, G. (eds.) COST 2102 Int. Training School 2010. LNCS, vol. 6456, pp. 147–154. Springer, Heidelberg (2011)
17. Gnisci, A., Di Conza, A., Zollo, P.: Political Journalism as a Democracy Watchman. In: Herrmann, P. (ed.) Democracy in Theory and Action, pp. 205–230. Nova Publisher, New York (2011)
18. Bull, P., Elliott, J., Palmer, D., Walker, L.: Why Politicians are Three-Faced: The Face Model of Political Interviews. Brit. J. Soc. Psychol. 35, 267–284 (1996)
19. Cruttenden, A.: Intonation. Cambridge University Press, New York (1986)
20. Fleiss, J.L.: Statistical Methods for Rates and Proportions. John Wiley & Sons, New York (1981)
21. Gnisci, A., Bonaiuto, M.: Grilling politicians: Politicians' Answers to Questions in Television Interviews and Legal Examinations. J. Lang. Soc. Psychol. 22, 384–413 (2003)
22. Gnisci, A., Di Conza, A., van Dalen, A., Graziano, E. Un Confronto tra Canali Televisivi Italiani nelle Ultime due Elezioni Politiche. In: X Congresso Nazionale della Sezione di Psicologia Sociale- Associazione Italiana di Psicologia, Torino, September 14-16 (2010) ISBN 978-88-905249-0-5
23. Gnisci, A., Graziano, E., Acerrano, V.: Si.Co.D.: Un Sistema Informatico per la Codifica delle Domande. Giornale Italiano di Psicologia 1, 175–190 (2012)
24. Cohen, J.A.: A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. 20, 37–46 (1960)

Rule-Based Morphological Tagger for an Inflectional Language

Daniel Hládek, Ján Staš, and Jozef Juhár

Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics, Technical University of Košice,
Park Komenského 13, 042 00 Košice, Slovak Republic
{daniel.hladek,jan.stas,jozef.juhar}@tuke.sk

Abstract. This paper aims to present an alternative view on the task of morphological tagging - a rule based system with new and simple learning method that uses just basic arithmetic operations to create an efficient knowledge base. Matching process of this rule-based approach follows specific-to-general technique, where rules for more specific contexts are applied whenever they are available in the rule-base. As a consequence, the major accuracy and performance improvements can be achieved by pruning the rule-base.

Keywords: Morphological tagger, rule-based, learning classifier system.

1 Introduction

Slovak language is characterized by a very rich morphology. When compared to the English language where only some inflections are allowed, one word can have many forms, according to its grammatical function. As a consequence, vocabulary of the Slovak language, that is a part of the Slavic language group, is much larger. Most of the current approaches to the morphological tagging are based on a statistics, calculated from a training corpus, where pairs (bigrams) or triplets (trigrams) of words are taken into the account. In this case is the vocabulary and a number of possible morphological tags much larger, thanks to the number of possible word forms. This makes the task of the morphological tagging more difficult. The word order in the language is non-mandatory and this fact means, that a morphological category of the word can depend on any surrounding context, history or following words. More on this topic can be found in [1,2].

To overcome this problem, it is possible to use cone of the traditional solutions. Current approaches in the part of speech tagging include hidden Markov models [3,4,5,6], and neural networks [7,8]. Each traditional statistical method that can be used for the task of morphological tagging is based on statistical information q, assigned to features f that describe word and its context. After learning weights q for every word, this information is then used to assign a result for each word, described by features f.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 208–215, 2012.

The problem is that in the practice is the whole learning process very complex. Usually it is hard to tell, how weights q in the knowledge base influence the final result. This problem is especially true for the case of the highly inflectional languages. Large vocabulary causes a high number of the statistical parameters.

On the other hand, in the rule-based techniques, knowledge is usually stored in a human readable form that can be easily processed and additional heuristics can be included more easily. This feature is very important, if a sufficient amount of manually annotated data is not available, such it is in the case of the Slovak language. The missing morphologically tagged data can be compensated by an expert knowledge, that can be inserted into the system in advance.

The proposed algorithm is inspired by the theory of the learning classifier systems [9] (a literature survey in [10,11]). Rules in the system are updated, created or destroyed according to their performance - contribution to the final solution. In the theory of the learning classifier systems, this process is viewed as a simulated evolution, where one rule in the system corresponds to an individual in an evolving population.

Expert knowledge can be utilized in several ways in this kind of the learning system. At first, it is easy to influence features that are taken into the account by the rule template. For each grammatical event that is important a special rule template can be used. Another way, natural to the rule-based system is a direct insertion or update of rules. If it is known that the automatically constructed rule-base of the system is wrong in some cases, it is possible to insert or update rules to correct the rule-base. Thanks to the easy readability of the rule it can be performed manually. The performance the system can also be influenced by human by adjusting its control constants.

2 Rule Structure

The basic part of the rule-based system is a rule, and in the proposed system it consists of:

- **Antecedent part** - it is a quintuple A that describes current word w and its context $\boldsymbol{c} \in C^w$, in this case it is $A = a_0, a_1, a_2, a_3, a_4$ where a_0, a_1 are preceding words, a_2 is the current word for which the resulting tag t' is searched and a_3, a_4 are following words in the context of the current word. Some parts of the antecedent string can be replaced with a wild-card to allow the rule to match more than one specific context. For example, antecedent part $A = (*, *, test, *, *)$ will match any context of the word *test*.
- **Consequent part** - is a list of all tags T that could be assigned to the antecedent of the rule.
- **Rule evaluation** - is a vector Q, where each attribute q_k corresponds to one of tags t_k in the consequent part of the rule.
  Total number of rule evaluations is defined as a total number of evaluations of the rule in the learning process, taking into account learning equation 4 it is $c_e = \sum_k q_k$.

3 Matching Process

The system contains a list of rules. This set of all rules has to be used to infer a final decision t'. When a word $w \in W$ with a certain context $\boldsymbol{c} \in C^w$ from a set of possible contexts of word w occurs, a **match set** M that contains a list of all rules that match word w and its context c is created. This set M is then used to create final decision. The outcome of the rule system decision is a certain tag t' that has to be chosen from the match set. After matching, match set M contains a list of rules whose antecedent parts A_k matches the word w and its current context $\boldsymbol{c}$. For every antecedent part A_k in rule r, there is a consequent part T_k with a list of tags $\boldsymbol{t}_k$ and their weights $\boldsymbol{q}_k$. Rules with $\boldsymbol{t}_k$ from the match set M forms a matrix with weights Q that is used to find the final decision t'. Value q' from Q marks result t'. The max-min strategy is used to obtain q' and t'.

First step of the decision is **maximization**. It means to choose the best tag t_k^{max} and its best weight q_k^{max} for every rule in the match set M. Maximization is performed for every line of matrix Q (for every matching rule) - a decision with maximal evaluation for each rule in the match set M is found:

$$t_k^{max} = arg \max_j q_{jk} \tag{1}$$

and

$$q_k^{max} = \max_j q_{jk}. \tag{2}$$

After maximization step, we have a list of tags t_k^{max} with their maximal weights q_k^{max} for each rule in the match set with antecedent part A_k that matches the current context.

Next step of the inference process is **minimization**. Values q_k^{max} of the matching rules are used to find tag t'.

Rules that are more specific and are evaluated enough (number of their evaluations is sufficient to deduce the correct consequent part), tend to have better results than more general rules. As a measure of generality of the rule, q_k^{max} value has been chosen. Rules with a larger number of correct guesses are taken as more general as those with smaller number of correct guesses. Therefore, rule with the smallest q_k^{max} that is the most specific, is used for the final decision. This step is performed by finding a tag with the minimal best weight q'.

$$t' = arg \min_k q_k^{max}. \tag{3}$$

The whole decision process then can be summarized as:

1. **Matching** - Creating match set M of rules that match word w and its current context $\boldsymbol{c}$.
2. **Maximization** - Finding tag t_k^{max} with maximal weight q_k^{max} for each rule k in match set M.
3. **Minimization** - Finding final tag t' as a tag with minimal weight q_k^{max}.

4 Learning Process

When designing a learning process of the rule system, a set of features that is used to determine result has to be composed. In the case of the morphological tagger, features that can be used are: a word and a portion of its context. These features create the rule template, that is used to create a new rule. According to the task, it is also possible to select different features, that are important and there is possibility, that will positively affect the result.

Features, that are taken into the account are:

- the word;
- the word and one preceding word;
- the word and one following word;
- the word and two preceding words;
- the word and two following words.

It is important to notice that these possible contexts can be partially ordered according to the covered possible contexts. The rule that is triggered just by matching single word is the most general and will be matched in all possible contexts. The rule that takes the word and the preceding word will cover a certain subset of all contexts covered by the most general rules. The rule with two preceding words will cover a subset of contexts covered by the rule with one preceding word. Rules with the following words can be compared to the rules with preceding words as almost equal in number of covered samples. This partial ordering according to the "generality" of the rule is fundamental to the algorithm.

The learning process is simple:

1. **Train sample taking** - When learning the system, a training sample as a word w, certain context $\boldsymbol{c}$ and correct tag t'' occurs.
2. **Matching** - First a list of matching rules is created. This means to find list of all rules in the match set that match the current word and its context.
3. **Rule creation** - After that, missing rules have to be added to the match set and the rule set, in order to have a complete match set.
4. **Rule update** - When the match set is complete, weights in the consequent part of the rule have to be updated. If the consequent part of the matching rule does not contain a correct consequent yet, it is simply added at the end. The update process in the match set is simple:
   For each rule in the match set, weight of the correct tag is increased by one. If the correct tag is not present in the consequent part of the rule, it is added. This very simple learning rule then can be written as:

$$q_c = q_c + 1, \tag{4}$$

   where c is the index of the correct tag. Values q of every other tag are left as they are. This step ensures that at the end of the learning there will be one tag that has captured the most correct hits and could be declared as the best matching tag for all samples covered by the rule.

5 Pruning Process

This learning process has several drawbacks:

- list of all possible contexts for one word is too high, thereafter the algorithm requires large amount of memory;
- if a number of evaluations of one rule is too low, resulting tag t' might be wrong;
- redundancy of information in the rule-base might be too high - a more specific rule can have the same result as a more general one, correct result for a certain word and a certain context is provided by more rules, some of rules are unnecessary.

To solve these issues, a pruning process is necessary. More general rules in the rule-base are possibly useful in more cases, but on the other hand, more specific rules can provide us better precision. Finding a good rule-base needs to choose a compromise between rule-base compactness and precision. In the presented approach a measure of rule generality is a number of evaluations $c_e = \sum_j q_j$ of all consequents of one rule.

The threshold-based pruning process starts with choosing a value $p(w)$ for every word in the rule-base that has been encountered in the learning process. Then every rule k that is matched by the word w and has a lower number of evaluations than $p(w)$,

$$c_{ek} \leq p(w)_k \tag{5}$$

is removed.

Value $p(w)$ is chosen, such that all rules k matched by word w are removed, if their number of evaluation c_{ek} is under average number of evaluations of all rules k matched by word w. For this heuristics, $p(w)$ is calculated as:

$$p(w) = p_c \frac{\sum_k c_{ek}}{\sum_k 1}, \tag{6}$$

where p_c is a constant from interval $(0, 1)$ that expresses how large portion of all rules k matched by word w should be pruned. Expression $\frac{\sum_k c_{ek}}{\sum_k 1}$ means average count of all rules with the same antecedent. Lower value of p_c means that pruning will remove less rules and the rule-base will be more specific. Higher values of p_c means that the pruning process removes more rules, leaving more general and more compact rule-base. This process is repeated once in a certain interval p_i during learning of the system. This value needs to be large enough, in order to allow rules to gather sufficient number of evaluations.

Exact size of the control constants p_c and p_i strongly depends on the problem solved by the system and size of the training data. More training data allows more evaluations of the rule in the system. Small number of p_c means more precise and bigger rule-base, smaller sizes mean more general and more compact rule-base.

Table 1. Testing corpora

	Words	Flags	Sentences
Corpus 1	524	2325	50
Corpus 2	463	1816	50

Table 2. Algorithm evaluation

	Corpus 1		Corpus 2	
	Errors	Corr [%]	Errors	Corr [%]
Rule-based POS tagger	113	78.43	19	95.89
HunPos POS tagger	104	80.15	27	94.16
Rule-based POS flags	137	94.08	17	99.06
HunPos POS flags	126	94.25	86	98.01

6 Evaluation

To evaluate the proposed algorithm, the SNK corpus [12,13] trigram counts with assigned morphological tags have been used to obtain the rule-base.

As a comparison, **HunPos** [14] morphological tagger (an implementation of TnT tagger [3]) with default parameters has been trained on the same trigram counts.

For testing, two testing corpora have been created. First, a corpus of 50 sentences from the Slovak parliament utterances. The second testing corpus is part of the Slovak classic novel, also 50 sentences long. Each sentence in the testing corpus has been manually checked for correctness. Testing corpora are characterized in the Table 1.

Control parameters prune threshold p_c has been set to 0.01 that assures that only rules with very small number of evaluations are removed, and prune process has been run once after $p_i = 6,000,000$ of items of the training dataset.

As a result, average number of rules per word was 47.5 with average rule count 106. These resulting counts depend on the pruning threshold parameters.

Results of evaluation are in the Table 2. As for testing corpus 1, reference tagger HunPos shows slightly better results (80.1% compared to 78.4%). In the testing corpus 2 situation changes, and the presented algorithm gets better (95.89% compared to 94.16%). Results in row 3 and row 4 in the Table 2 display accuracy when evaluating individual flags. Each morphological flag consists of several individual flags for every grammatical category that is taken into the account.

The presented values have to be taken just as a reference, because accuracy values are also influenced by the accuracy of the training corpus of the Slovak language that have been partially tagged using statistical tagger. Anyway, according to the presented results it is possible to say, that the performance of the presented approach is at least as good as performance of current state-of-the art, hidden Markov-model based part-of-speech tagger.

7 Conclusion

Importance of this work is in providing new, alternative view on a topic that has been studied for a long time. The presented algorithm can be used for other tasks such as sentence segmentation or named entity resolution, where disambiguation of meaning is required.

Thanks to its rule-based approach it is easily possible to modify the process of creation of the rule-base and choose features that will describe the observed event in the best way. In this point, also custom feature-extraction function can be used as a part of the rule template.

There is still room for accuracy and speed improvement in enhancing of the pruning process. By choosing a value of the pruning threshold, it is possible to influence the size and compactness of the rule-base. The small rule-base can be very quick and easy to process, on the other hand bigger rule base can assure higher precision. Anyway, these features should be the target of the future work.

Acknowledgement. The research presented in this paper was supported by the Ministry of Education under the research project MŠ SR 3928/2010-11 (50%) and Research and Development Operational Program funded by the ERDF under the project ITMS-26220220141 (50%).

References

1. Nouza, J., Zdansky, J., Cerva, P., Silovsky, J.: Challenges in Speech Processing of Slavic Languages (Case Studies in Speech Recognition of Czech and Slovak). In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) COST 2102 Int. Training School 2009. LNCS, vol. 5967, pp. 225–241. Springer, Heidelberg (2010)
2. Beňuš, S., Cerňak, M., Rusko, M., Trnka, M., Darjaa, S.: Adapting slovak asr for native germans speaking slovak. In: Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, DIALECTS 2011, pp. 60–64. Association for Computational Linguistics, Stroudsburg (2011)
3. Brants, T.: Tnt: A statistical part-of-speech tagger. In: Proc. of the Sixth Conference on Applied Natural Language Processing, ANLC 2000, pp. 224–231. Association for Computational Linguistics, Stroudsburg (2000)
4. Collins, M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 1–8. Association for Computational Linguistics (2002)
5. Schmid, H., Laws, F.: Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1, pp. 777–784. Association for Computational Linguistics (2008)
6. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2003, vol. 1, pp. 173–180. Association for Computational Linguistics, Stroudsburg (2003)

7. Spoustová, D., Hajič, J., Raab, J., Spousta, M.: Semi-supervised training for the averaged perceptron pos tagger. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2009, pp. 763–771. Association for Computational Linguistics, Stroudsburg (2009)
8. Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P.: The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In: Proc. of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies, pp. 67–74. Association for Computational Linguistics (2007)
9. Holland, J.H.: Escaping brittleness: the possibilities of general-purpose learning algorithms applied to parallel rule-based systems. Machine Learning: An Artificial Intelligence Approach 2 (1986)
10. Sigaud, O., Wilson, S.: Learning classifier systems: a survey. Soft Computing-A Fusion of Foundations, Methodologies and Applications 11(11), 1065–1078 (2007)
11. Hládek, D.: Learning System Based on Generalization of Fuzzy Rules. PhD thesis, Technical University of Kosice (2009)
12. Jazykovedný ústav Ľ. Štúra SAV: Slovenský národný korpus prim-3.0-public-all (2007)
13. Horák, A., Gianitsová, L., Šimková, M., Šmotlák, M., Garabík, R.: Slovak National Corpus. In: Sojka, P., Kopeček, I., Pala, K., et al. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 89–93. Springer, Heidelberg (2004)
14. Halácsy, P., Kornai, A., Oravecz, C.: HunPos - An open source trigram tagger. In: Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL 2007, pp. 209–212. Association for Computational Linguistics, Stroudsburg (2007)

Czech Emotional Prosody in the Mirror of Speech Synthesis

Jana Vlčková-Mejvaldová[1,2] and Petr Horák[1]

[1] Institute of Photonics and Electronics, Academy of Sciences of the Czech Republic
[2] Faculty of Education, Charles University in Prague
{vlckova,horak}@ufe.cz

Abstract. Contemporary speech synthesisers still provide a fairly monotonous and tedious output when used for longer Czech texts. One of the ways how to make these texts more lively is the synthesis of emotionally coloured speech. In the present paper we focus on the modelling of real-speech-based emotions in synthetic speech and the subsequent assessment of emotionally coloured utterances in listening tests with the aim of determining the role that individual prosodic parameters play in the identification of each emotion.

Keywords: speech synthesis, prosody, emotions, Czech.

1 Introduction

Speech contains linguistic information conveyed by verbal, lexical or prosodic means. The latter category includes information about sentence modality, permitting the differentiation of declaratives and questions.

Besides linguistic information, speech also inevitably contains paralinguistic and extralinguistic information, which can be found mainly in the prosodic component. The border between these two categories is not always distinct; they include information about the age, sex, regional and social origin of the speaker, as well as his or her current physical and psychological condition. Last but not least, the sound structure of the utterance provides information about the attitude of the speaker towards the given communicative situation and its components.

Emotions affect the quality of individual segments (i.e. speech sounds) through variations of articulatory strength. Listeners can intuitively discriminate articulations produced under the effect of anger (especially when it is held in by the speaker), with clenched teeth etc., and articulations of a bored speaker, which is extremely lax, with highly reduced muscular activity. Apart from articulation, the muscle tonus also affects suprasegmental (i.e. prosodic) characteristics of the utterance.

Speech synthesis is not only a valuable tool for the seeing-impaired, but is also used in everyday situations by people without any handicap, e.g. in train station announcements, voice navigation, automatic information systems over the telephone etc. Attempts have been made to use synthetic speech e.g. for book reading.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 216–224, 2012.

2 Aims of the Present Work

Global improvements of synthetic speech quality, i.e. higher naturalness, better comfort in longer listening and higher variability of the sound structure, require taking into consideration those properties of natural speech which are responsible for the aforementioned phenomena.

In speech synthesis, which operates with an invariable sound inventory, the simulation of emotions is limited to prosody. The quality of sound segments, be they coded in diphones, triphones or other units, is constant, whereas the prosodic form of the utterance can be modulated by means of proper adjustments of prosodic parameters [1].

To ensure that a synthetic utterance is perceived as more natural, and to "make it supportable for a longer time", it is necessary that there is a certain prosodic variability within each emotion [2]. In our current research, we have focused on the impact of each of the three prosodic parameters (i.e. F0, duration and intensity) on the identification of emotions as expressed by synthetic prosody.

Our approach not only leads to an improvement of the quality and the naturalness of synthetic speech, but enables us to indirectly investigate natural speech as well; as a matter of fact, modifications of a single parameter in natural utterances are in principle impossible.

3 Modelling Emotional Prosody in Synthetic Speech

In the process of modelling emotional prosody, we had to respect specific properties of synthetic speech when adjusting the parameters, because exaggerating the changes would have led to a decrease in the naturalness of the utterance.

In the different versions of each synthetic emotion, we only changed either one parameter at a time (F0, duration and intensity, respectively), or a combination of these. Since most of the studies have been concerned with the influence of F0 on the perception of marked and unmarked synthetic sentences, but have not provided fully satisfying results, we have extended the scope of our study to the role of the other two parameters in the identification and the convincingness of the emotion expressed by synthetic speech [3].

For applying the values of individual prosodic parameters from the natural onto the synthetised speech, we modelled the sentence Předpověď počasí na zítřek slibuje vysoké teploty a bezvětří. ['pr̝ɛtpɔvjɛc 'pɔʃasi: 'nazi:tr̝ɛk 'slɪbujɛ 'vɪsɔkɛ: 'tɛplɔtɪ ʔa'bɛzvjɛtr̝i:] 'The weather forecast for tomorrow indicates high temperature and no wind'.

3.1 Synthesis type

The model sentences were created by means of the "Epos" TTS system [4, 5], using triphone synthesis in the time domain (PSOLA method). The prosody used in the model sentences was the one which is defined as intrinsic in the Epos TTS system, i.e.

automatically generated, rule-based neutral prosody. We used the “machac” speech unit inventory (sampling frequency 16 kHz) for the male voice and the “violka” inventory (sampling frequency 32 kHz) for the female voice [6]. The model sentences were produced by means of the “Winsay” program, an extension of the “Epos” TTS, intended for speech-unit-based modelling of prosody [7]. The prosody automatically generated by the present system is in principle limited to F0 changes; the intensity (I) of speech units remains constant throughout the utterance, and no variability in the time (T) domain is implemented either.

4 Marked Prosody in Natural and Synthetic Speech

Naturally enough, we based the modelling of emotionally coloured sentences, conveying extralinguistic information, on expressive prosody of natural speech. The model sentence chosen was *Mně bylo samozřejmě vod samýho začátku prostě zcela jasný, že to nemůže dopadnout jinak* ['mɲɛ'bɪlɔ 'samozr̝ejmɲɛ 'vɔtsamɪ:hɔ 'zat͡ʃa:tku 'prɔscɛ 'st͡sɛla'jasnɪ:ɬʒɛ tɔ 'nɛmu:ʒɛ 'dɔpadnou̯t 'jɪnak] “Of course, I knew from the very beginning that it would end up this way”, which was taken from a longer text, written by the speaker himself (professional actor, 54 years old). Like this, the speaker had an opportunity to choose an adequate stylisation. Using the same sentence avoid different syntactic influences on the sentence prosody, which allow us to compare most clear emotional prosody, without an extra/emotional influences.

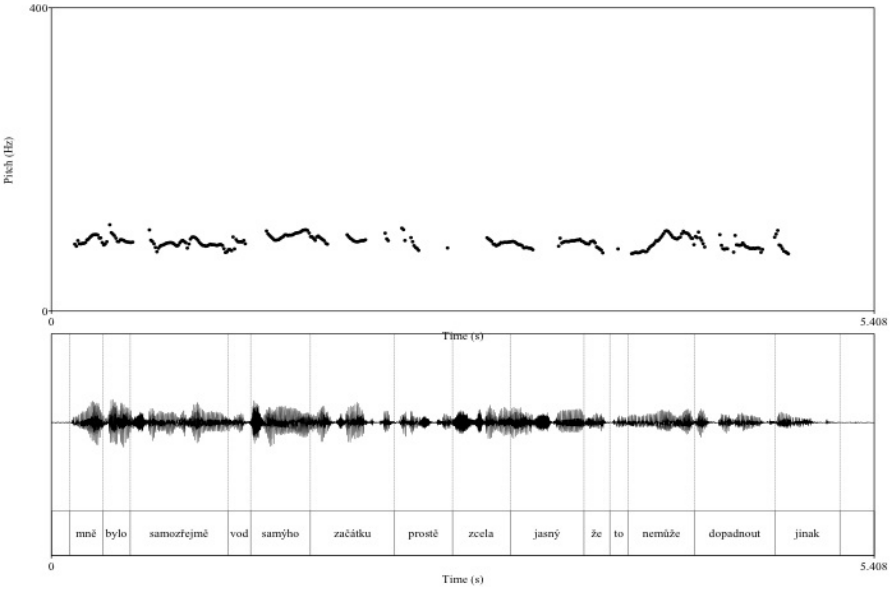


Fig. 1. Prosody of an unmarked natural utterance

For further experiments, we chose four emotions – joy, anger, boredom and sadness. Prosodic characteristics which are typical of each of the aforementioned emotions [8] were applied to the synthetic utterance by means of manual modelling. The labelling was realised by means of the Praat software [9]. Prosodic features of the tested sentences are displayed in this software as well.

In the next part of our paper, we would like to present the prosodic form of experimentally obtained sentences expressing different emotions. We were concerned with the prosody of joy, anger, sadness and boredom. For the sake of comparison, we first give the neutral realisation of the sentence as a reference.

As far as the experimental recording is concerned, the following characteristics turned out to be typical of the individual emotions:

4.1 Joy

This emotion is classified as positive and active. It is characterised by greater melodiousness (i.e. intonational variability), as well as higher average F0, which reaches 174.4 Hz, while it does not exceed 90 Hz in the neutral version produced by the same speaker. Greater variability can also be observed in the temporal domain: the articulation rate is higher, while some syllables are lengthened, so that the overall duration of the utterance is 6.5 s (vs. 5.1 s in the neutral version); higher duration also affects the pause. These findings were the basis for our modelling of individual sentences supposed to express this emotion.

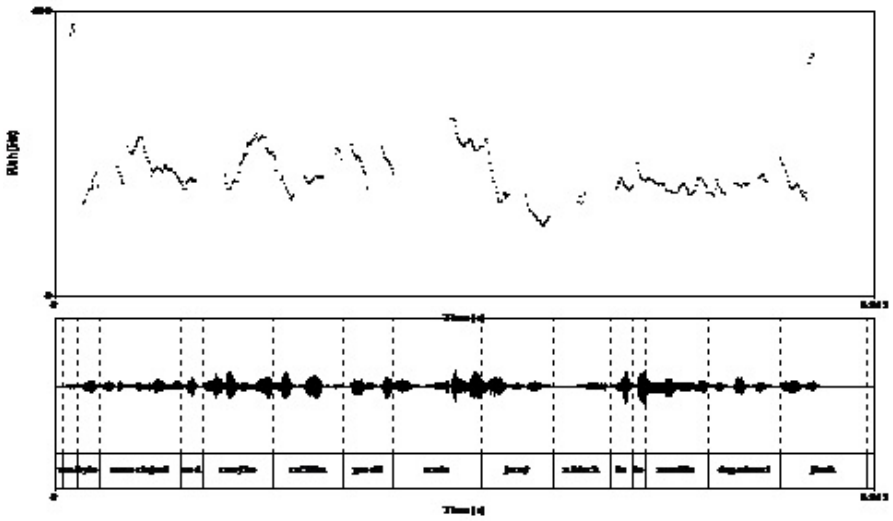


Fig. 2. Prosodic expression of joy in a natural utterance

When applying time proportions of the natural sentences to speech synthesis, we changed the duration of individual segments from the automatically generated value (75%) to 65% in stressed syllables, and to 55% in unstressed syllables. We also sought to achieve a higher variability in the dynamic profile of the sentences: the

segment intensity was increased up to 250% in stressed syllables, while it fell to 75% in the final part of the sentence. As far as fundamental frequency is concerned, our strategy was again to achieve high variability.

4.2 Anger

To characterise utterances expressing this kind of emotion in prosodic terms, we can say that they exhibit strong stressing and both dynamic and melodic variability. Consonants are lengthened, stops have longer occlusions.

As for temporal structure, we synthesised this emotion by concentrating the most salient changes in the final part of the sentence; we avoided the standard lengthening which affects units at the end of neutral declarative sentences. Besides that, segments in stressed syllables are twice as long as those in unstressed syllables (up to 140% against 70%, respectively). There are marked stresses, the stressed syllables achieving up to 270% of the intensity value. The F0 course is characterised by a slight fall on the stressed syllable, followed by a rise on the post-tonic syllable, and another fall on the subsequent syllables of the stress group, if there are any. The overall falling melodic trend is maintained.

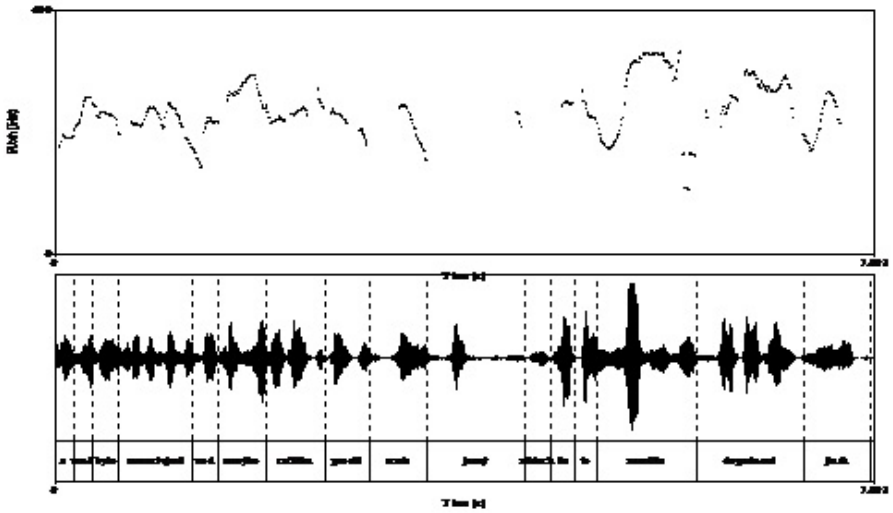


Fig. 3. An utterance with prosodic marks of anger

4.3 Boredom

Slow tempo, low F0 variability (if any), a generally low F0 level and a reduced intensity are the prosodic phenomena which accompany this emotion.

The intonation is monotonous, average F0 level is around 95 Hz. The global F0 trendline is clearly falling. Phonetic characteristics of boredom also include slow articulation rate, lengthening of final syllables in syntactic groups, audible breath intakes and, naturally, extremely lax articulation.

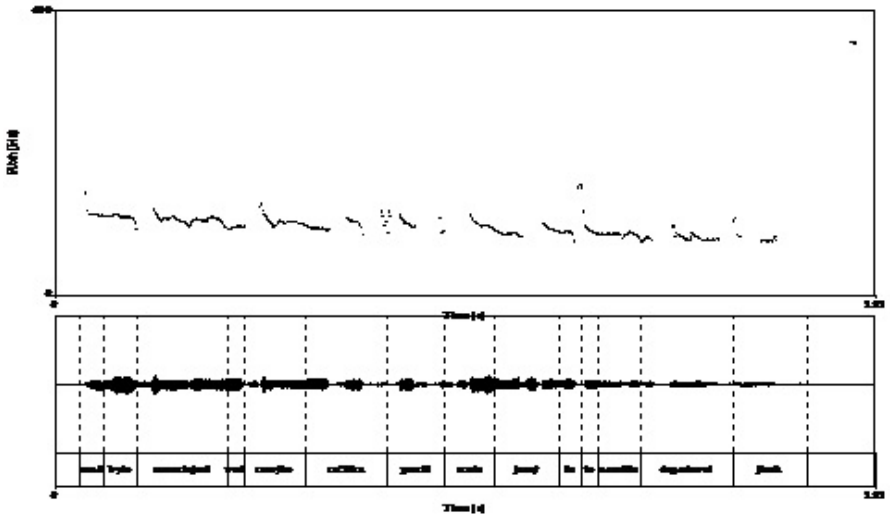


Fig. 4. Prosodic realisation of boredom

The latter factor plays an important role in the identification of this emotion in natural speech. However, changing the quality of individual segments in this respect is not possible in the synthesis system. When modelling the sentences, we increased the duration of individual segments up to 160% of the automatically generated value in stressed syllables, and up to 110% in unstressed syllables. Final segments were lengthened up to a maximum of 170%. It is worth noting that the overall F0 level is lower when compared to that of sadness: the sentence starts at 87% of the automatically generated value, exhibits tiny variations in stressed syllables (90%) and then gradually falls to 83% in the final part of the utterance. The intensity is slightly higher than that used in sadness: it varies between 90% in stressed syllables and 60% at the end of the sentence.

4.4 Sadness

As far as prosody is concerned, this emotion is very similar to boredom, except for average F0, which is slightly higher (128 Hz in the analysed utterance), and exhibits tiny melodic variations. The intensity is reduced, and the speech rate slow.

The segment duration was increased up to 115–150% in stressed syllables, while it was not lower than 100% in unstressed syllables (cf. the 75% value in automatically generated neutral prosody). The intensity varies between 50% in unstressed syllables and 75% in stressed syllables, in contrast to automatically generated neutral prosody, which – as it has been mentioned above – only works with the 100% value. As regards the melodic pattern of the manually modelled sentences included in the test, the F0 starts at 98% and very slowly falls to 93%. With respect to the sentence length, these changes are virtually negligible.

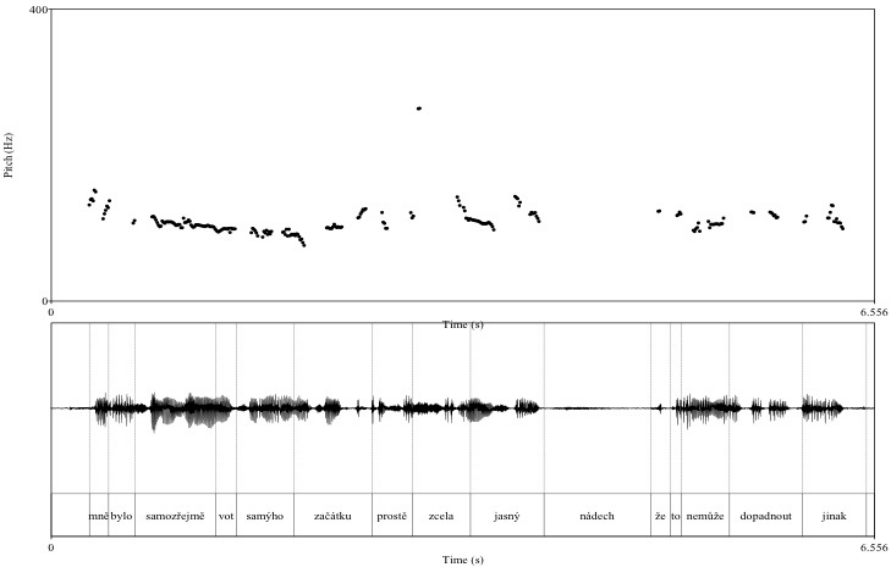


Fig. 5. An utterance with prosodic marks of sadness

5 Results of Perception Tests

Simple synthetic sentences with neutral semantics were modified prosodically. A set of such modified sentences was subject to perception tests. The hypothesis is that a correct identification of different emotions from natural, but also synthetic sentences (the latter being the object of our study) is based on relevant changes of different parameters.

To make the perception tests simpler and to ensure that their results can be interpreted unambiguously, we used a single sentence which was synthesized in a male voice. Four different emotions (anger, sadness, joy and boredom) were prosodically implemented by means of manual (but automatable) adjustments of one or more prosodic parameters (F0, I, T, F0+I, F0+T, I+T, F0+I+T). The following graph shows how relevant the individual prosodic parameters are for the identification of emotions in synthetic speech. One can notice that for each of the four examined emotions, the best results are achieved with different means: F0 changes either isolated or in combination with the other two parameters, are relevant for the identification of joy. Anger was best identified on the basis of F0 and I (or F0 and T) changes; isolated changes of intensity also lead to satisfactory results. The identification of sadness is closely connected with F0 changes, which is the most important parameter; surprisingly, combined changes of intensity and time also yield good identification of this emotion. Duration is by far the most relevant parameter for the identification of boredom.

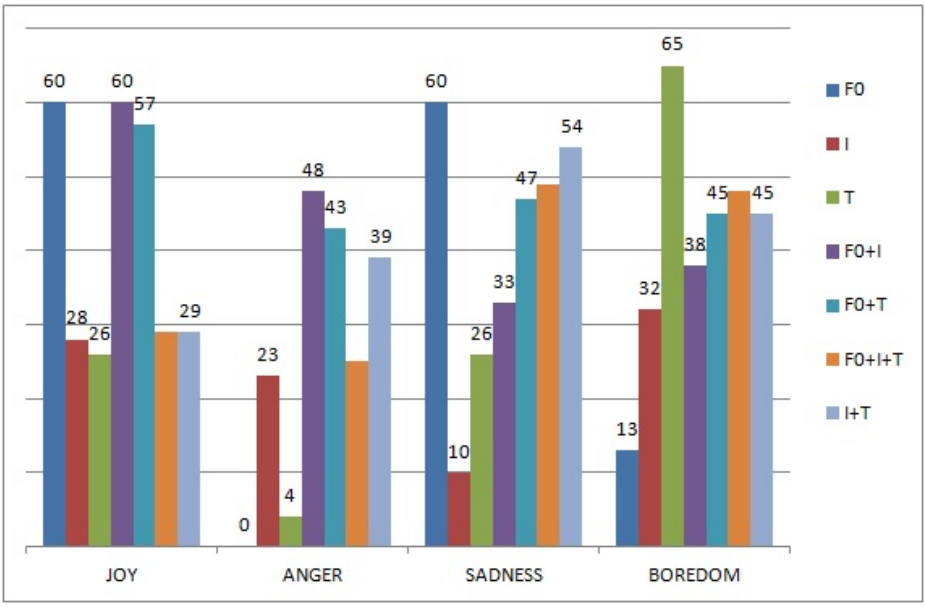


Fig. 6. Correct identification rate for individual emotions as based on the changes of different prosodic parameters

6 Conclusion

The experimental approach used in the present study confirmed the hypothesis that the identification of different emotions is based on relevant changes of different parameters and their combinations. For different emotions different prosodic parameter is responsible of its correct identification. It can even become that changing more than one parameter is, for the identification of the emotions, misleading, what was shown by the results of perception tests.

It is obvious that the description of a concrete synthetic pattern, whether it led to a correct identification of the emotion, or to any type of mismatch or wrong interpretation, is tied to a concrete language, i.e. Czech in the present case. However, it is possible to generalise the initial hypothesis according to which the identification of different emotions is based on relevant changes of different prosodic parameters (i.e. F0 and/or duration and/or intensity). Our findings can be used for enhancing the prosodic variability of synthetic speech, a feature required especially in longer synthesized texts or in texts whose contents is not exclusively informative (e.g. synthesising short pieces of fiction by means of automatic emotion tagging based on text analysis).

Acknowledgement. This research was realised with the support of the GA ČR 102/09/0989 grant project.

References

1. Chaloupka, Z., Horák, P.: Prosody Modelling Possibilities of the Czech Emotional Speech. In: Proceedings of 19th Czech-German Workshop Speech Processing, Prague, pp. 114–117 (2009)
2. Vlčková-Mejvaldová, J., Horák, P.: Prosodic Parameters of Emotional Synthetic Speech in Czech: Perception Validation. In: Travieso-González, C.M., Alonso-Hernández, J.B. (eds.) NOLISP 2011. LNCS (LNAI), vol. 7015, pp. 170–176. Springer, Heidelberg (2011)
3. Dohalská, M., Mejvaldová, J., Duběda, T.: Prosodic Parameters of Synthetic Czech: Can We Manage without Duration and Intensity? In: Keller, E., Bailly, G., et al. (eds.) Improvements in Speech Synthesis, pp. 29–133. Wiley & Sons, Chichester (2001)
4. Hanika, J., Horák, P.: Epos – A New Approach to the Speech Synthesis. In: Proceedings of the First Workshop on Text, Speech and Dialogue – TSD 1998, Brno, Czech Republic, September 23-26, pp. 51–54 (1998)
5. Hanika, J., Horák, P.: Dependences and Independences of Text-to-Speech. In: Palková, Z., Wodarz, H.-W. (eds.) Forum Phoneticum 70. Frankfurt am Main, pp. 27–40. Hector Verlag (2000)
6. Epos system documentation, `http://epos.ufe.cz/`
7. Horák, P., Hesounová, A.: Czech Triphone Synthesis of Female Voice. In: Speech processing. In: Proceedings of 11th Czech-German Workshop, Prague, pp. 32–33 (2001)
8. Vlčková-Mejvaldová, J.: Prozodie, cesta a mříž porozumění. Praha, Karolinum (2006)
9. Praat software, `http://www.praat.org`

Pre-attention Cues for Person Detection

Karel Paleček[1], David Gerónimo[2], and Frédéric Lerasle[3,4]

[1] Institute of Information Technology and Electronics,
Technical University of Liberec, Czech Republic
[2] Computer Vision Center, Autonomous University of Barcelona, Spain
[3] CNRS: LAAS, 7 Avenue Colonel Roche F-31077 Toulouse, France
[4] Université de Toulouse, UPS, INSA, INP, ISAE, LAAS-CNRS, Toulouse, France
karel.palecek@tul.cz, dgeronimo@cvc.uab.es, lerasle@laas.fr

Abstract. Current state-of-the-art person detectors have been proven reliable and achieve very good detection rates. However, the performance is often far from real time, which limits their use to low resolution images only. In this paper, we deal with candidate window generation problem for person detection, i.e. we want to reduce the computational complexity of a person detector by reducing the number of regions that has to be evaluated. We base our work on Alexe's paper [1], which introduced several pre-attention cues for generic object detection. We evaluate these cues in the context of person detection and show that their performance degrades rapidly for scenes containing multiple objects of interest such as pictures from urban environment. We extend this set by new cues, which better suits our class-specific task. The cues are designed to be simple and efficient, so that they can be used in the pre-attention phase of a more complex sliding window based person detector.

Keywords: person detection, candidate window generation, pre-attention.

1 Introduction

In the last two decades, human detection has been an active research area of computer vision. The algorithms for human detection are applicable in various tasks, e.g. surveillance systems [15], driver assistance [7,11] or human-machine interaction [10].

However, fast and robust human detection is still a challenging task for several reasons. The main problem is large variability of appearance, i.e. people can wear different clothes or take various poses. Typically, there are also other common factors such as variability in camera poses, lighting conditions or overall image quality.

Although several approaches for human detection in still images have been adopted over the years, some of the most successful detectors are based on a sliding window technique. This class of algorithms treats human detection task as a classification problem, i.e. input image is sequentially scanned and each sub-window is classified as containing or not containing a human. Since the size of the human figure is not known a priori, this procedure is repeated for different

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 225–235, 2012.

scales. State-of-the-art examples of such classifiers are Histogram of Oriented Gradients (HoG) [6] or part-based models [8]. These classifiers are based on low level gradient features, similar to ones introduced in [14]. They are computed over small blocks of 16×16 pixels and classified using linear Support Vector Machine (SVM).

However, due to complexity of gradient features and large number of windows that needs to be evaluated, one of the main problems of these classifiers is their computational complexity. For example, processing a 640×425 grayscale image with Felzenswalb's detector [8] takes roughly 7 seconds on a 3 GHz Core Duo machine with 8 GB RAM.

We aim our work at reducing the search space that needs to be evaluated by sliding window based classifiers and therefore speeding-up the detection process. In other words, we deal with the problem of candidate window generation for person detection, i.e. we want to discard as many false positive windows and keep as many true positive ones as possible by utlizing less costly algorithms than the ones used during the final classification.

We base our work on the paper on candidate window generation by Alexe et al. [1], which deals with generic object detection. Alexe et al. propose a framework with a set of four cues and a window sampling procedure, which serves as a preprocessing step for the robust state-of-the-art classifiers. Cues of the framework are designed such that they favor regions that are likely to contain an object of any class. They claim their framework to speed-up several state-of-the-art object detectors [6,8] by up to 20-40 times by reducing the total number of windows that needs to be evaluated. In this paper, we evalute this framework in the context of person detection, modify if and extend it by considering another set of cues, specifically suited for person detection task.

The structure of this report is as follows. Sect. 2 describes the pre-attention cues and their fusion, Sect. 3 then evaluates and discusses achieved results. Finally, in Sect. 4 we present conclusions and propose future extensions.

2 Description of the Pre-attention Cues

We first review the cues proposed in [1] and then we propose additional cues suited for person detection task.

2.1 Objectness Cues

Multiscale Saliency. Multiscale Saliency (MS) is based on spectral residual approach of Hou et al. [12], which has high response for regions that are unique in terms of appearance within an image. Typical examples of such regions are object on uniform background. The image f is first down-sampled to some predefined size $s_0 \times s_0$. The saliency map annotated $I^{s_0}_{MS}$ is computed by inverse FFT of the residual of original and smoothed log-spectrums of the image. In order to extract objects at different scales MS repeats this procedure for several image

sizes s and for each of them a saliency map I^s_{MS} is obtained. MS score of window r is then computed as

$$MS(r, \theta^s_{MS}) = \sum_{\mathcal{P}} I^s_{MS}(p) \times \frac{|p \in r \mid I^s_{MS}(p) \geq \theta_s|}{|r|} \tag{1}$$

where θ^s_{MS} are free threshold parameters for each of the scale s and $\mathcal{P} = \{p \in r \mid I^s_{MS}(p) \geq \theta_s\}$. In order to extract regions of different sizes, MS is computed for every $m' \times n'$ sub-window r of the resulting saliency maps, where $m', n' = 1, \ldots, s$.

Color Contrast. Color Contrast (CC) measures color dissimilarity between a region and its surroundings. Surroundings $\mathrm{Surr}(r, k_{CC})$ of a window r is defined as a rectangular ring obtained by scaling the window r proportionally by factor k_{CC} in and subtracting the area of window r. The dissimilarity is computed as Chi-square distance of color histograms of the window and its surroundings, i.e.

$$CC(r, k_{CC}) = \chi^2 \left(h(r), h\left(\mathrm{Surr}(r, k_{CC})\right)\right) \tag{2}$$

To compute the histograms $h(r)$ and $h\left(\mathrm{Surr}(r, k_{CC})\right)$ of the window r and its surroundings $\mathrm{Surr}(r, k_{CC})$, image is converted to Lab space and then quantized into l color levels. Note that with relatively small number of quantized colors the histograms can be computed in constant time by a table look-up using the summed area tables (integral images) trick.

Edge Density. Edge Density (ED) captures the fact that images of objects usually have well defined borders while they do not have many edge pixels inside. The ED is computed as a density of edge pixels near the window borders, i.e.

$$ED(r, k_{ED}) = \frac{\sum_{p \in \mathrm{Inn}(r, k_{ED})} I_{ED}(p)}{\mathrm{Len}(\mathrm{Inn}(r, k_{ED}))}, \tag{3}$$

where $\mathrm{Inn}(r, k_{ED})$ is the inner ring of window r obtained by shrinking it by factor k_{ED} (similarly to CC) and $\mathrm{Len}(\cdot)$ is its perimeter. Edge pixels are obtained by Canny edge detector [3].

Superpixels Straddling. Similarly to ED, Superpixels Straddling (SS) cue captures the closed boundary characteristics of an object. It is based on superpixel segmentation [9], which segments the image into small regions of uniform color. After segmentation, surfaces of objects consist of several superpixels, which preserve their boundaries. The SS measures the extent to which the superpixels straddle the test window r. A superpixel is straddling a window r if it contains at least one pixel inside and one pixel outside r. The degree, by which a superpixel straddles window r, is defined as the minimum of the number of its pixels inside r and the number of its pixels outside r. SS score of window is then computed as a sum of degrees of straddling for all superpixels contained in r, i.e.

$$SS(r, \theta_{SS}) = 1 - \sum_{s \in Sl(\theta_{SS})} \frac{\min\left(|s \backslash r|, |s \cap r|\right)}{|r|}, \tag{4}$$

where θ_{SS} is a segmentation scale.

2.2 Proposed Cues

Since in our task we want to find candidate windows which might contain persons rather than just generic objects, we might take advantage of some specific features. As persons in images from video surveillance or driver assistance systems are usually in upright positions, we mainly try to explore the vertical symmetry and edge properties of candidate regions.

Color Symmetry. Color Symmetry (CS) cue is based on comparing the color distributions of the inner and outer parts of a test window. Ideally, the windows containing persons should be vertically symmetrical in terms of color, i.e. the left and right part of the bounding box should have similar color distribution, whereas the person itself should be distinctive from its local neighborhood. Each window r of size $m \times n$ is divided into four parts in the direction of x-axis: left half $r_L(r)$, right half $r_R(r)$, inner rectangle $r_I(r)$ and outer part $r_O(r)$. The inner rectangle has size $m \times n/2$ pixels and is positioned in the center of the window r. The outer rectangle covers the rest of the window r. The CS score is then computed as

$$CS(r) = \frac{\chi^2\left(h_I(r), h_O(r)\right)}{\chi^2\left(h_L(r), h_R(r)\right) + \epsilon}, \tag{5}$$

where $\chi^2\left(h_x(r), h_y(r)\right)$ is the Chi-squared distance of color histogram of the rectangles $r_x(r)$ and $r_y(r)$ and ϵ is a smoothing parameter. In order to compute the histograms efficiently and avoid additional computation demands, we use the same approach for quantization of the colors as in the case of Color Contrast cue, that is the image is first converted into Lab color space and then quantized into l color levels.

Edge Symmetry. Edge Symmetry (ES) is another symmetry-based cue. Similarly to [16], it exploits the fact that person often appears in an image as either bright or dark blob, which is roughly symmetrical. As mentioned in [2], the left and right parts of the window should contain similar amount of edges, but their sign in the direction of x-axis should be opposite. We therefore vertically divide the window r into k rectangular areas $r_{1\ldots i\ldots k}(r)$ of size $\left\lfloor \frac{n}{k} \right\rfloor \times m$ and on each of them we compute the total sum $s_i^L(r)$ and $s_i^R(r)$ of their left and right edge pixels. We call an edge pixel $p = (x, y)$ left if $I(x-1, y) < I(x+1, y)$, i.e. it lies on a transition from dark to bright region. ES score of window r is computed as

$$ES(r) = \sum_{i=1}^{k} \left(s_i^L(r) - \overline{s^L(r)}\right) \cdot \left(s_{k-i+1}^R(r) - \overline{s^R(r)}\right), \tag{6}$$

We subtract the mean values $\overline{s^L(r)}$ and $\overline{s^R(r)}$ from $s_i^L(r)$ and $s_i^R(r)$ in order to suppress the score of windows with many randomly distributed edges. Similarly to ED, edges are obtained using Canny edge detector.

Verticality. Usually images of persons contain much more vertical edges than edges of other directions. Verticality (VE) tries to use this property by computing the relative amount of vertically oriented edge pixels in a window r. It is similar to Edge orientation histogram descriptor (EOH) [13], but it differs in several aspects. The number of orientation bins b_i is fixed to 4, for better robustness the edges are detected using Canny edge detector rather than simple thresholding of the convolution of the image with Sobel kernel and it considers only the vertical orientation bin value. It is computed as

$$VE(r) = \frac{E_3(r)}{\sum_i E_i(r)}, \tag{7}$$

where

$$E_i(r) = \sum_{p \in r} \psi_i(p) \tag{8}$$

are sums of edge pixels p, which have orientation ψ_i, i.e.

$$\psi_i(p) = \begin{cases} 1 & \text{if} \theta(p) \in b_i \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Thus, $E_3(r)$ corresponds to the vertical bin. Note that similarly to the CC and CS cues the values of $E_i(r)$ can be computed rapidly by computing summed area table for each of the four orientations.

Dominant Orientation. Since VE considers only the value of the vertically oriented bin, the natural question that arises is whether it would be beneficial to utilize the values of the other bins too. We construct a normalized window descriptor $E^n(r) = [E_1^n, \ldots, E_4^n]^\top$, where

$$E_i^n = \frac{E_i(r)}{\sum_j E_j(r)}, \tag{10}$$

and classify the window using linear classifier w_{DO}. The score is then computed as

$$DO(r) = w^\top E(r) + \beta \tag{11}$$

In the case of VE score computation 7 the classifier corresponds to $w = [0, 0, 1, 0]^\top$ and $\beta = 0$, i.e. only the relative value of the vertically oriented bin E_3 8 is considered. In DO 2.2 the coefficients of the classifier w and β are found by training a linear SVM [5] on a set of positive and negative window samples. We use libSVM [4] for SVM classification.

2.3 Cue Combination

In order to take advantage of all information available, cues can be combined into a single classifier. The score of each cue is computed for every position of the sliding window, which has a fixed size. Different scales are processed by recursively reducing the size of the image. Note that this approach is different from that in [1], where the cues are only computed for windows sampled from MS score distribution.

Naïve Bayesian classifier is used to compute the final score of a test window. The score $p(\mathrm{p}|\mathcal{A})$ of set of cues $\mathcal{A}$ is computed as

$$p\left(\mathrm{p}|\mathcal{A}\right) = \frac{p\left(\mathrm{p}\right)\prod_{c\in\mathcal{A}} p_c\left(x_c|\mathrm{p}\right)}{p\left(\mathrm{p}\right)\prod_{c\in\mathcal{A}} p_c\left(x_c|\mathrm{p}\right) + p\left(\bar{\mathrm{p}}\right)\prod_{c\in\mathcal{A}} p_c\left(x_c|\bar{\mathrm{p}}\right)}, \tag{12}$$

where $p(\mathrm{p})$ and $p(\bar{\mathrm{p}})$ are prior probabilities of finding a person and background, respectively, $p_c\left(x_c|\mathrm{p}\right)$ is the probability of the score x_c respective to the score distribution $p_c\left(\mathrm{p}\right)$ of the cue c. Rather than assuming a continuous probability density such as normal distribution, the probability distributions p_{p} and $p_{\bar{\mathrm{p}}}$ are modeled as histograms, which are computed over a training dataset.

3 Experiments

We evaluate the cues on INRIA person dataset[1] as it is standardized dataset containing persons in different contexts, scales and image quality. The free parameters of the cues were trained on the training subset of the INRIA person database containing 2416 images. The test subset comprised of 288 annotated images containing total of 589 persons (number of persons in one image varies from 1 to 16).

Fig. 1. Example images from INRIA person dataset

[1] http://pascal.inrialpes.fr/data/human/

3.1 Evaluation of Cues

Cues are evaluated based on the pyramid scheme. Each image is first proportionally resized to the maximum allowed size of 640×480 pixels for efficiency reasons. Then, an image pyramid is built by recursively reducing the size of the image using bilinear interpolation. We use $L = 18$ pyramid levels with scale factor $\kappa = 2^{-1/6}$. Search window size is fixed to 20×50 pixels for all pyramid layers.

For every cue true positive ratio (TPR) vs. false positive ratio (FPR) curve (ROC) is plotted, see Fig. 2 and Fig. 3. We consider a test window positive if it is covering any of the ground truth windows of the image. The window r is covering the window o if the Pascal criterion

$$PC(r, o) = |r \cap o| \,/\, |r \cup o| \tag{13}$$

is larger than some threshold t_{PC}, i.e. $PC(r, o) > t_{PC}$. In our case, we use $t_{PC} = 0.7$. Thus, TPR is a ratio of correctly classified positive windows and the total number of all positive windows. Note that using this definition of positive and negative samples has two advantages. The first advantage is the ability to generate nearly arbitrary number of training samples. The second advantage is that the training samples are similar to the input in the testing phase, i.e. due to sparse sampling, the sliding window is often poorly aligned to ideal position over an image of person.

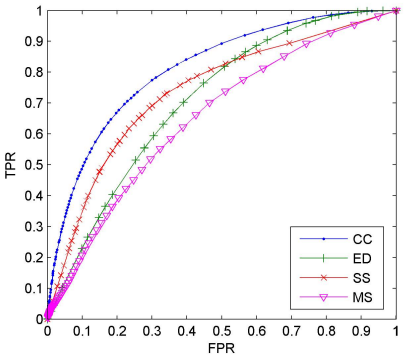


Fig. 2. ROC of the cues proposed in [1]

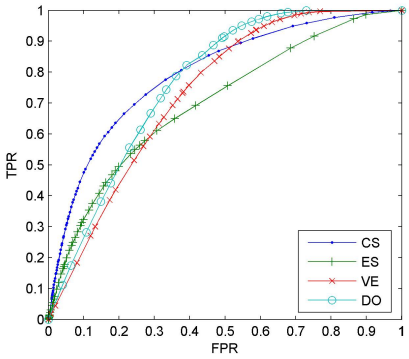
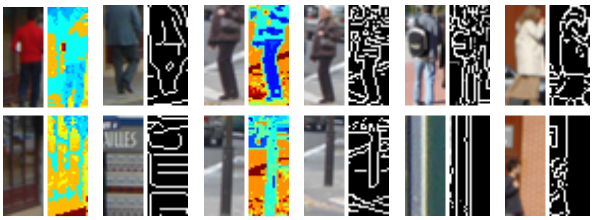


Fig. 3. ROC of our proposed cues

The obtained results for individual cues are shown in Tab. 1. For each cue, the FPR is false positive ratio when 95% of true positive windows are preserved.

The edge-based cues, mainly VE and DO, performed very well in our tests. Their performance depends on the ability to extract edges from an image reliably. They are both global features in terms of the test window, which makes them robust to noise and missing edge information. Also, they are both very fast to evaluate, since they can be computed using the summed area table trick [17].



(a) (b) (c) (d) (e) (f)

Fig. 4. True positive (top) and false positive (bottom) examples: a) CC, b) ED, c) CS, d) ES, e) VE, f) DO

Symmetry-based cues work best for test windows that are precisely aligned to some ground truth window. Both ES and CS are sensitive to misalignments of the test and ground truth windows. This can be overcome by using finer scale-factor κ in order to cover the ground truth windows by more percent in terms of the Pascal criterion. However, the obvious problem with this approach is an increase of computational costs.

As opposed to [1], SS cue did not perform best in our test. The main problem of the superpixel segmentation is its sensitivity to blur and other image quality degradations, which results in violation of the key assumption of SS that superpixels preserve object boundaries. While decreasing the value of segmentation scale θ_{SS} increases the overall recognition rate, it also increases the computational complexity because of the large number of extracted regions that needs to be evaluated for every window.

We also found that the Multiscale Saliency is not well suited for detecting multiple objects within a single image. This is due to its spectral residual approach, which only favors regions with unique appearance, not repetitive patterns.

The examples of false and true positive windows as classified by the cues are shown in Fig. 4.

Table 1. Results for Pascal criterion threshold 0.7 and TPR = 0.95

Cue	MS	CC	ED	SS	CS	ES	VE	DO
FPR	0.89	0.66	0.72	0.85	0.69	0.82	0.61	0.55

3.2 Cue Combination

We evaluated all 127 possible cue combinations. The results for selected combinations as well as for original objectness measure from [1] are shown in Fig. 5. One of the best result was obtained using only three cues: Color Contrast, Color Symmetry and Dominant Orientation. The achieved false positive rate at 0.95 true positive rate was 0.33. The false positive rate at the same true positive rate

for general objectness measure was 0.50. When the computationally inefficient SS cue was not taken into account, the objectness false positive rate was 0.63, almost twice as large as the best case. The improvement of our extended set over the general objectness measure is caused by two reasons: utilization of person specific charateritics and poor performance of Multiscale Saliency for cluttered scenes or images with multiple persons.

Adding another cues did not lead to significantly better performance in our experiments. This can be explained by the fact that cues that are based on the same type of information (e.g. color) are quite correlated. As one can expect, the most correlated pair of cues with normalized correlation coefficient $\rho = 0.76$ are VE and DO, which differ only in classification of the edge histogram.

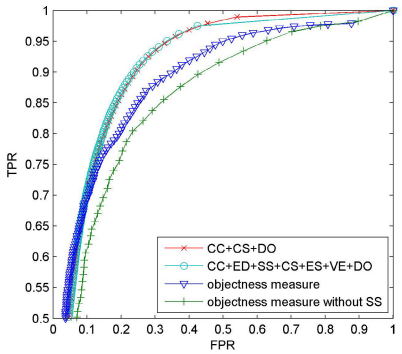


Fig. 5. ROC curves for selected cue combinations

3.3 Time Consumption

We also evaluated time consumption of the cues. All experiments were performed on PC with Core 2 Duo @ 3GHz processor and 8 GB RAM. We implemented the algorithms in Matlab and C++ using OpenCV[2] library for image processing tasks and the HoG detector. We use Felnzenswalb's code[3] for the segmentation for SS cue. Since each group of cues (e.g. color-based) performs similar preprocessing steps such as color quantization or edge detecting, we measure the time cost of the preprocessing and the window score computation separately. The results are shown in Tab. 2 and Tab. 3.

Longer computation times of color-based compared to edge-based cues are caused by relatively large number of integral histograms, which have to be computed for every quantized color. Similar problem causes extremely long computation times for SS cue, where the integral histogram is computed for each superpixel. Depending on the segmentation scale θ_{SS}, there can be up to hundreds of superpixels in a cluttered scene. We also evaluated our candidate window

[2] http://opencv.willowgarage.com/wiki/
[3] http://www.cs.brown.edu/~pff/segment/

Table 2. Preprocessing

Color-based cues	331 ms
Edge-based cues	46 ms
Superpixels	7617 ms

Table 3. Score computation

MS	35 ms	CS	461 ms
CC	640 ms	ES	113 ms
ED	133 ms	VE	82 ms
SS	11255 ms	DO	82 ms

generation algorithm together with the HoG detector [6]. With exhaustive search on a dense grid, 83% detection rate was obtained with the HoG detector, while processing the cca 0.8 MPix images took 46 seconds on average. When using the pre-attention phase with CC, CS and DO cues and thresholding 12, the overall achieved detection rate was 69% with the average of 4 seconds per image. In other words, we achieved a speed-up factor of 11.5 while preserving 84% of the true positives.

4 Conclusion

We have evaluated several pre-attention cues for person detection that can be used to reduce the search space of arbitrary sliding window based detector. We have shown that cues proposed in [1] do not perform well in the task of person detection, especially in the cases where image contains cluttered background, multiple objects of interest or blur. In order to solve these problems, we have proposed additional cues specifically suited for person detection. Also, Tab. 1 shows better efficiency of the proposed cues. In our experiments, cues utilizing edge orientation properties achieved the lowest false positive rate. They also are more efficient than color-based cues, which rely on computing color histograms. Symmetry-based cues performed well only when combined with other type of features. Superpixel Straddling cue did not perform well in our experiments, both in false positive rates and efficiency. As a next development we will investigate how to optimally combine the cues into more efficient classifier, possibly using multiple stages of classification.

Acknowledgments. The research reported in this paper was partly supported by Student Grant Scheme at Technical University of Liberec.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), pp. 73–80 (2010)
2. Bertozzi, M., Broggi, A., Del Rose, M., Felisa, M.: A symmetry-based validator and refinement system for pedestrian detection in far infrared images. In: Intelligent Transportation Systems Conference, pp. 155–160. IEEE (2007)
3. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8, 679–698 (1986)

4. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995), doi:10.1007/BF00994018
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893. IEEE Computer Society (2005)
7. Enzweiler, M., Gavrila, D.M.: Monocular Pedestrian Detection: Survey and Experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(12), 2179–2195 (2009)
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vision 59, 167–181 (2004)
10. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots (2003)
11. Gerónimo, D., López, A.M., Sappa, A.D., Graf, T.: Survey of Pedestrian Detection for Advanced Driver Assistance Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(7), 1239–1258 (2010)
12. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8 (June 2007)
13. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004), pp. 53–60. IEEE Computer Society (2004)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
15. Ogale, N.A.: A survey of techniques for human detection from video. Survey, University of Maryland (2006)
16. Schauland, S., Kummert, A., Park, S.B., Iurgel, U., Zhang, Y.: Vision-based pedestrian detection – improvement and verification of feature extraction methods and svm-based classification. In: Intelligent Transportation Systems Conference, ITSC 2006, pp. 97–102. IEEE (September 2006)
17. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: CVPR, pp. 511–518 (2001)

Comparison of Complementary Spectral Features of Emotional Speech for German, Czech, and Slovak

Jiří Přibil[1] and Anna Přibilová[2]

[1] Institute of Measurement Science, SAS, Dúbravská cesta 9, SK-841 04 Bratislava, Slovakia
Jiri.Pribil@savba.sk
[2] Institute of Electronics and Photonics, Faculty of Electrical Engineering & Information Technology, Slovak University of Technology, Ilkovičova 3, SK-812 19 Bratislava, Slovakia
Anna.Pribilova@stuba.sk

Abstract. Our paper is aimed at statistical analysis and comparison of spectral features which complement vocal tract characteristics (spectral centroid, spectral flatness measure, Shannon entropy, Rényi entropy, etc.) in emotional and neutral speech of male and female voice. This experiment was realized using the German speech database EmoDB and the Czech and Slovak speech material extracted from the stories performed by professional actors. Analysis of complementary spectral features (basic and extended statistical parameters and histograms of spectral features distribution) for all three languages confirms that this approach can be used for classification of emotional speech types.

Keywords: spectral features of speech, emotional speech, statistical analysis.

1 Introduction

Identification of emotions in speech depends on the chosen set of features extracted from the speech signal. These features are systematically divided into segmental and supra-segmental ones [1]. Short-term segmental features derived from speech frames with short duration are usually in relation with the speech spectrum. These include traditional features like linear predictive coefficients, line spectral frequencies, mel-frequency cepstral coefficients, linear prediction cepstral coefficients [2], or unconventional ones like perceptual linear predictive coefficients, log frequency power coefficients [3], mel bank spectrum features [4], or spectrally weighted mel-frequency cepstral coefficients [5]. Supra-segmental features comprise statistical values of parameters describing prosody by duration, fundamental frequency, and energy. Included in this category is also a separate group of features constituting the voice quality parameters: jitter, shimmer, glottal-to-noise excitation ratio, Hammarberg index [6], normalized amplitude quotient, spectral tilt, and spectral balance [2]. Several spectral features (spectral centroid, spectral flatness measure, Shannon entropy, Rényi entropy, etc.) are used to complement the mentioned segmental and supra-segmental features for speaker recognition [7]. Investigation must be done whether these features bear also information about various emotions manifested in speech together with information about various speakers.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 236–250, 2012.

This paper describes analysis and comparison of complementary spectral features (CSF) of male and female acted speech in four emotional states: joy, sadness, anger, and a neutral state. Motivation of our work was to find out whether the complementary spectral features determined from emotional speech depend on the speaker nationality. From our previous research follows that using the speech material uttered of Czech and Slovak speakers, the CSF values depend only on the speaker (including the gender) and the emotional style of the performed speech. As the Czech and Slovak languages are very similar, this hypothesis need not be valid in the cases of speech spoken in other languages. Therefore we perform analysis of CSF values obtained from the German speech database EmoDB [8] described completely in [9] and the Czech and Slovak speech material extracted from the stories performed by professional actors [10], [11]. If our hypothesis is correct then the statistical distribution of CSF values will be different for neutral and emotional speech (divided into two classes by the gender – male/female voice) for all three languages. Otherwise, statistical similarities will be found which cause incorrect classification of analysed CSF values to the corresponding emotion group. This work originated from our previous research of complementary spectral features in Czech and Slovak [12]. Results will be used together with values of the basic spectral properties and prosodic parameters for creation of the training data corpus for the emotional speech classifier based on statistical approach (like Gaussian mixture models used for speaker recognition and identification [13]) that is currently being developed.

2 Subject and Method

Our experiments are aimed at statistical analysis and comparison of the CSF in emotional and neutral speech. It comprises comparison of basic statistical parameters (minimum, maximum, mean values, and standard deviation) and calculated histograms of distribution. Extended statistical parameters (skewness, kurtosis) are subsequently calculated from these histograms and/or the histograms can be evaluated by the analysis of variances (ANOVA) approach. Hypothesis tests are used for objective classification of neutral and different emotional styles. This statistical approach is often applied also in other research areas [14], [15].

2.1 Definition of Complementary Spectral Features

Complementary spectral features can be determined during cepstral speech analysis (see left part of the block diagram in Fig. 1) using the absolute value of the fast Fourier transform $|S(k)|$ of the speech signal $x(n)$ and the spectral power density $P(k)$

$$S(k) = \sum_{n=1}^{N_{FFT}} x(n) \; e^{-j\frac{2\pi}{N_{FFT}}nk}, \quad P(k) = \frac{|S(k)|^2}{\sum_{k=1}^{N_{FFT}/2} |S(k)|^2}, \tag{1}$$

where N_{FFT} represents the number of processed points for FFT calculation.

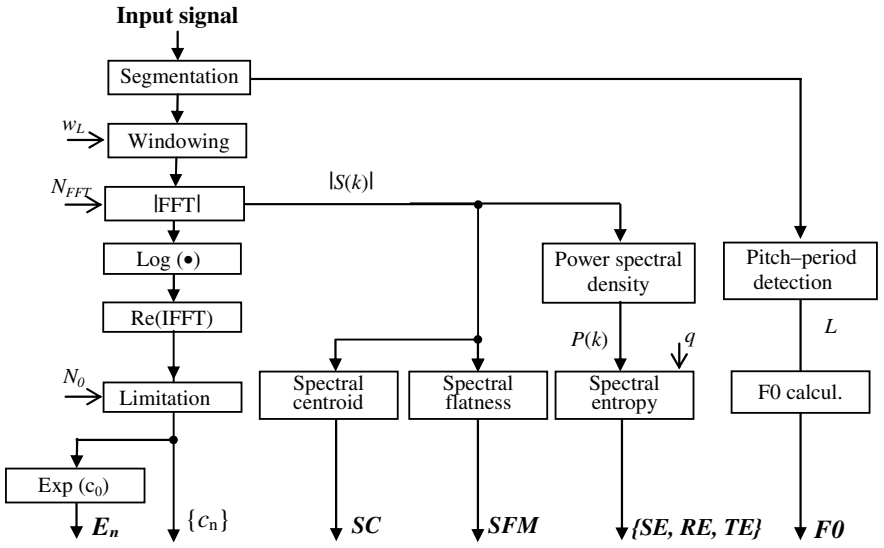


Fig. 1. Block diagram for calculation of complementary spectral features of the speech signal

The spectral centroid (SC) is a centre of gravity of the power spectrum [7]. It is an average frequency weighted by the values of the normalized energy of each frequency component in the spectrum. It is a measure of spectral shape and "brightness" of the spectrum (higher centroid values correspond to "brighter" voice with more high frequencies). The SC in [Hz] can be calculated as

$$SC = \frac{\sum_{k=1}^{N_{FFT}/2} k\left|S(k)\right|^2}{\sum_{k=1}^{N_{FFT}/2} \left|S(k)\right|^2} \cdot \frac{f_s}{N_{FFT}}, \tag{2}$$

where f_s is the sampling frequency.

According to psychological research of emotional speech different emotions are accompanied by different spectral noise. In cepstral speech synthesis the spectral flatness measure (SFM) was used to determine voiced/unvoiced energy ratio in voiced speech analysis [16], and also the voicing transition frequency for the harmonic speech model [10] determined by this parameter. The SFM values lie generally in the range of (0÷1) – the zero value represents the signal that is totally voiced (for example pure sinusoidal signal); in the case of SFM = 1, the totally unvoiced signal is classified (for example white the noise signal). This spectral feature can be calculated by the following formula

$$SFM = \frac{\left[\prod_{k=1}^{N_{FFT}/2} \left|S(k)\right|^2\right]^{\frac{2}{N_{FFT}}}}{\frac{2}{N_{FFT}} \sum_{k=1}^{N_{FFT}/2} \left|S(k)\right|^2}. \tag{3}$$

Spectral entropy is a measure of spectral distribution [17], [18]. It quantifies a degree of randomness of spectral probability density represented by normalized frequency components of the spectrum. The structured speech has lower entropy; the non-structured speech has higher entropy. Spectral entropy will be low for spectra having clear formants whereas for unvoiced sounds it will be higher. Shannon spectral entropy (SE) is defined as

$$SE = -\sum_{k=1}^{N_{FFT}/2} P(k)\log_2 P(k). \tag{4}$$

Generalizations of Shannon spectral entropy are:

– Rényi spectral entropy (RE)

$$RE = \frac{1}{1-q}\log_2 \sum_{k=1}^{N_{FFT}/2} P(k)^q, \tag{5}$$

where q is the order of the entropy.

– Tsallis spectral entropy (TE)

$$TE = \frac{1}{q-1}\sum_{k=1}^{N_{FFT}/2}\left[P(k) - P(k)^q\right]. \tag{6}$$

They are more sensitive to small changes in spectrum because of the exponent term q. For $q \to 1$, they reduce to Shannon entropy. Different applications use different orders of RE an TE; e.g. RE of the order 1.1 and 1.9 was examined in [19], 2nd and 3rd order RE was used in [20], RE of the order 3.5 and TE of the order 2 were examined in [21] – see examples in Fig. 2.

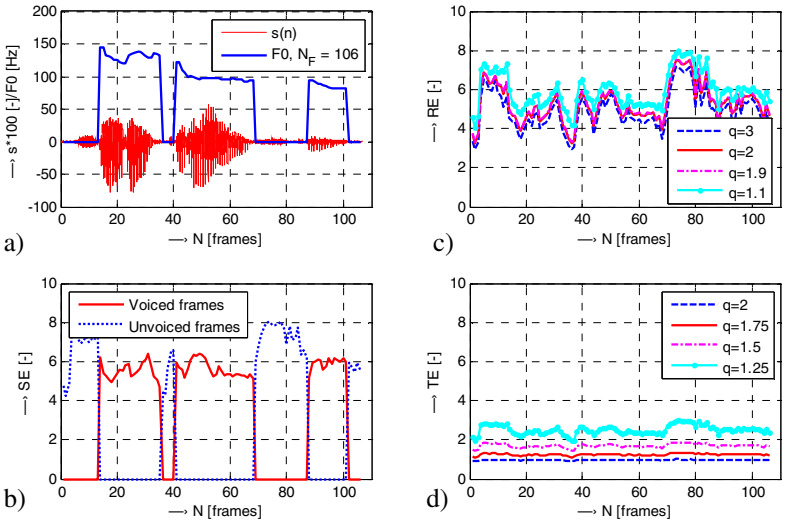


Fig. 2. Example of spectral entropy determination – sentence "Dcera královská" (King's daughter), Czech male voice, f_s = 16 kHz: speech signal together with F0 contour (a), SE contour in dependence of the speech signal voiceness (b), RE contours for q={3, 2, 1.9, 1.1} (c), TE contours for q={2, 1.75, 1.5, 1.25} (d).

2.2 Calculation of Complementary Spectral Features

Calculation of CSF values is supplied with determination of the fundamental frequency F0 and the energy En contour (calculated from the first cepstral coefficient c_0) – see Fig. 1. For voiceness frame classification, the value of the detected pitch-period L was used. If the value $L \neq 0$, the processed speech frame is determined as voiced, in the case of $L = 0$ the frame is marked as unvoiced. On the border between voiced and unvoiced part of the speech signal a situation can occur when the frame is classified as voiced, but the CSF value corresponds to the unvoiced class. For correction of this effect, the output values of the pitch-period detector are filtered by a 3-point recursive median filter.

In our algorithm, the values of SC and SFM are obtained only from the voiced speech frames, and in the case of spectral entropies (Shannon, Rényi, and Tsallis) from voiced and unvoiced frames with the signal energy higher than the threshold $\mathrm{En}_{\min}$ – for elimination of speech pauses between words within the sentence and beginning and ending parts of the sentence (see example in Fig. 3).

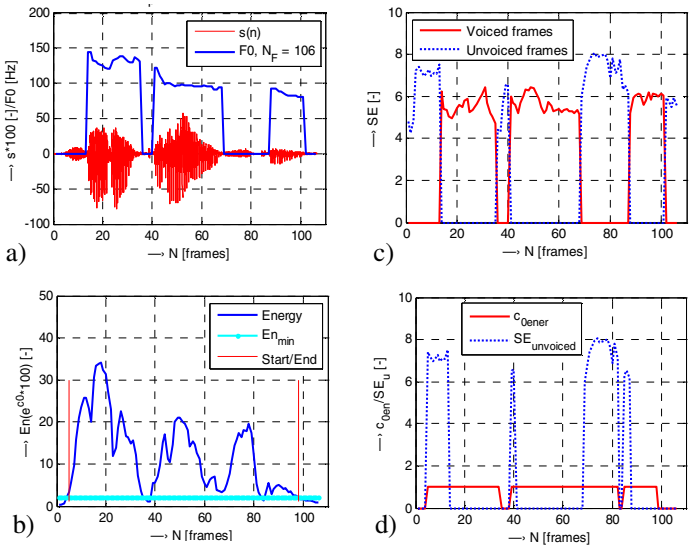


Fig. 3. Example of sentences processing: speech signal together with F0 contour (a), En contour calculated from the first cepstral coefficient c_0, determined threshold $\mathrm{En}_{\min} = 0.02$, and eliminated beginning and ending parts (b), calculated SE contour with determination of voiced and unvoiced frames (c), clipping function by $\mathrm{En}_{\min}$ threshold applied to SE unvoiced frames (d) – processed sentence is the same as in Fig. 2.

Parameters SC and SFM exhibit great differences between values determined from voiced and unvoiced speech signal; therefore only voiced frames of speech signal were analyzed. It is not valid in the case of the spectral entropy; hence all speech frames must be analyzed here. Because of similar statistical results of the Rényi entropy, Tsallis entropy, and Shannon entropy and sensitivity to small changes in the spectrum due to the exponent term q (see demonstration example in Fig. 4 and numerical comparison of values in Table 1), only values of Shannon entropy were finally evaluated and compared.

Obtained CSF values are processed separately in dependence on the voice type (male / female), subsequently sorted by emotional styles, and stored in separate stacks – see block diagram in Fig. 5. The whole process of statistical analysis of CSF values can be divided into seven steps:

1. determination of the mean values and basic statistical parameters,
2. calculation and building of histograms,
3. calculation of extended statistical parameters from histograms,
4. visual comparison of calculated histograms,
5. calculation of the mean value ratios of the CSF for emotional and neutral states,
6. application of the ANOVA supplemented with multiple comparison of group means,
7. numerical matching by the hypothesis test.

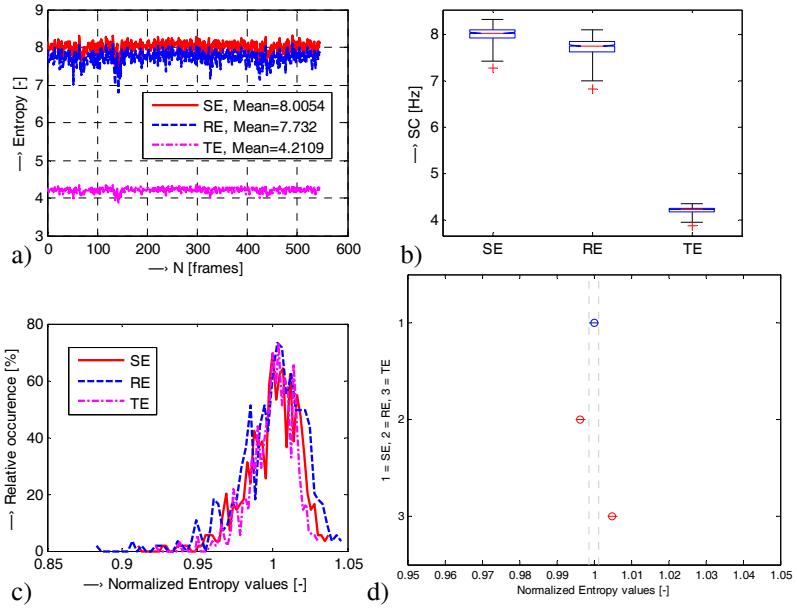


Fig. 4. Example of statistical similarity of three types of spectral entropy: obtained values of SE, RE ($q = 1.25$), and TE ($q = 1.1$) of testing noise signal (a), box plot of basic statistical parameters (b), normalized histograms (c), difference between group means with the help of ANOVA statistics (d)

Table 1. Compariosn of basic and extended statistical parametrers of spectral entropy values obtained frm the testing noise signal (see corresponding graphs in Fig. 4)

Value type	SE	RE [-]				TE [-]			
	[-]	q=3	q=2	q=1.9	q=1.25	q=2	q=1.75	q=1.5	q=1.1
Mean	8.01	6.79	7.20	7.26	7.73	0.99	1.30	1.85	4.21
Std	0.15	0.30	0.24	0.24	0.17	0.001	0.004	0.01	0.06
Skewness *)	-0.97	-1.18	-1.29	-1.29	-1.13	-2.48	-2.04	-1.64	-1.09
Kurtosis *)	1.83	2.65	3.35	3.37	2.62	11.08	7.84	5.29	2.38

*) calculated from normalized histograms

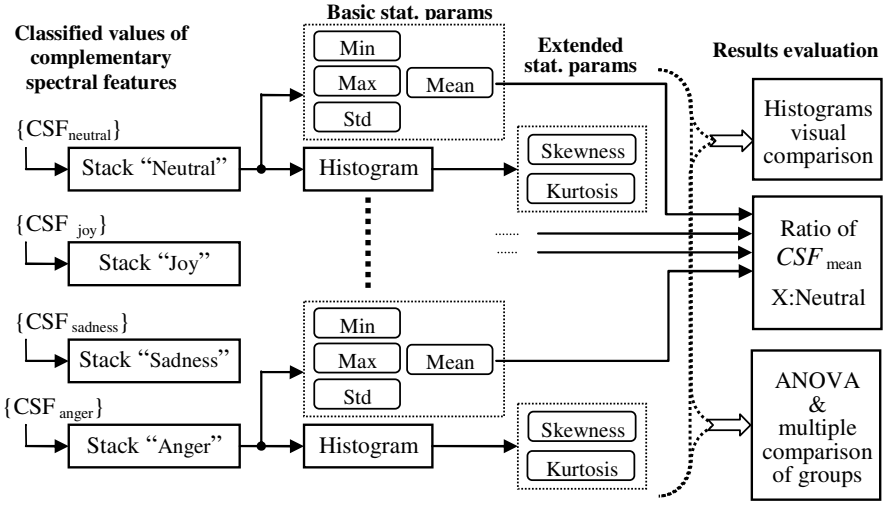


Fig. 5. Block diagram of CSF values collective processing for different emotional states

3 Material, Experiments, and Results

The complementary spectral features depend on a speaker as well as the emotions of a speaker. In our experiment the speech material from two databases was analyzed and compared. The first speech corpus ("A") was taken from the Berlin Database of Emotional Speech EmoDB [8], [9] in German language. This speech database consists of a set of sentences with the same contents expressed in seven emotional styles: "neutral", "joy", "sadness", "boredom", "fear", "resistance", and "anger". For our comparison we use only four emotional types - "neutral", "joy", "sadness", and "anger". We extracted 95 sentences spoken by 5 male speakers, and 134 sentences spoken by 5 female speakers with duration from 1.5 to 8.5 seconds (f_s = 16 kHz). The second speech corpus ("B") was extracted from the Czech and Slovak stories performed by professional actors and contains sentences with different contents expressed in four emotional styles: "neutral", "joy", "sadness", and "anger" uttered by several speakers (134 sentences spoken by male voices, and 132 sentences spoken by female voices, 8 + 8 speakers altogether). Processed speech material consists of sentences with duration from 0.5 to 5.5 seconds, resampled at 16 kHz.

The F0 values (pitch contours) were given by autocorrelation analysis method [22] with experimentally chosen pitch-ranges by visual comparison of testing sentences (one typical sentence from each of the emotions and the voice classes) as follows: 55 ÷ 250 Hz for male, and 105 ÷ 350 Hz for female voices. Then, the F0 values were compared and corrected by the results obtained with the help of the PRAAT program [23] with similar internal settings of F0 values. The frame length depends on the mean pitch period of the processed signal. Since the speech material collected in both databases (for male / female voices) originates from speakers (S_mA / S_fA, S_mB / S_fB) with different mean F0 value (see Table 2), different parameter settings for speech signal analysis – frame (window) length L_W and window overlapping L_O must be applied. Therefore three classes of input parameters were determined (C1–3_m and

C1−3$_f$) for both speech corpora. Current used window lengths for speech signal analysis are shown in Table 3. The energy threshold En_{min} calculated from the first cepstral coefficient c_0 was experimentally set to 0.015 for all processed sentences.

Table 2. Speaker mean F0 values – male and female voice of both speech corpora

Speaker	S_{m1}	S_{m2}	S_{m3}	S_{m4}	S_{m5}	S_{m6}	S_{m7}	S_{m8}	S_{f1}	S_{f2}	S_{f3}	S_{f4}	S_{f5}	S_{f6}	S_{f7}	S_{f8}
$F0_{mean}$ [Hz] A)	126	102	107	136	118	–	–	–	201	196	224	175	225	–	–	–
$F0_{mean}$ [Hz] B)	133	127	98	132	99	111	108	88	208	229	177	207	197	200	185	211

A) Emo DB, B) CZ & SK stories

Table 3. Input parameters for speech signal segmentation and analysis

Speaker class *)	$C1_m$	$C2_m$	$C3_m$	$C1_f$	$C2_f$	$C3_f$
L_W [samples]	256	328	164	128	180	164
Speakers from corpus "A"	S_{m3}, S_{m5}	S_{m2}	S_{m1}, S_{m4},	S_{f3}, S_{f5}	S_{f2}, S_{f4}	S_{f1}
Speakers from corpus "B"	S_{m6}, S_{m7}	S_{m3}, S_{m5}, S_{m8}	S_{m1}, S_{m2}, S_{m4}	S_{f2}	S_{f3}, S_{f5}, S_{f7}	S_{f1}, S_{f4}, S_{f8}

*) $L_O = L_W/2$, $N_{FFT} = 1024$, $f_s = 16$ kHz

Speech signal analysis of the corpus "A" was performed for the total number of 16234 frames (male speakers), and 25753 frames (female speakers). The SC and SFM values were determined from the voiced frames, altogether 12964 spoken by male speakers and 19458 spoken by female speakers. In the case of the speech material from the corpus "B", the total number of the analyzed frames was 25988 for male speakers, and of 24017 frames for female speakers; the SC and SFM values were determined from the voiced frames, altogether 11693 spoken by male speakers, and 13464 spoken by female speakers.

Obtained results of CSF values and their statistical evaluations are structured to sets corresponding to the type of the analyzed feature: spectral centroid, spectral flatness measure, and Shannon spectral entropy. It means that every set comprises of results for different speech styles, separately for male and female voices obtained by processing of sentences from both analyzed speech corpora. The results are presented in graphical as well as numerical form:

- Box-plot graphs of basic statistical parameters – see Figures 6 – 8a, b, c, d).
- Graphs of filtered histograms – see Figures 6 – 8e, f, g, h).
- Graphs with visualization of the differences between group means calculated using ANOVA statistics – see Figures 6 – 8i, j, k, l).
- Tables with corresponding resulting null hypothesis/probability values for 5% significance level of the Ansari-Bradley test – see Tables 4-6.

Results of basic statistical analysis of CSF values (the mean values and standard deviations) are presented in Table 7, obtained results of extended statistical analysis – skewness and kurtosis parameters determined from histograms for male and female voices in neutral and emotional states – are stored in Table 8. The summary results – CSF value ratios between different emotional states and a neutral state for male and female voices – are given in Table 9.

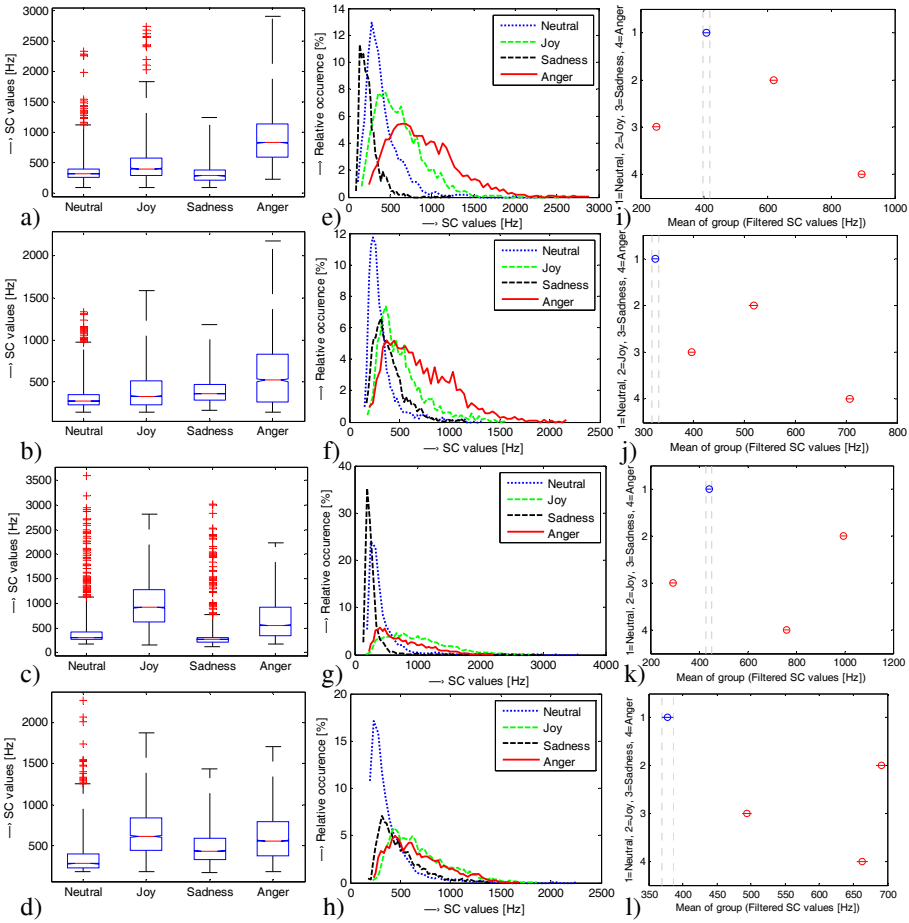


Fig. 6. Results of SC values – boxplot of basic statistical parameters: corpus "A" male voice (a), corpus "A" female voice (b), corpus "B" male voice (c), corpus "B" female voice (d); histograms: corpus "A" male voice (e), corpus "A" female voice (f), corpus "B" male voice (g), corpus "B" female voice (h); difference between group means with the help of ANOVA statistics: corpus "A" male voice (i), corpus "A" female voice (j), corpus "B" male voice (k), corpus "B" female voice (l); determined from voiced frames only.

Table 4. Null hypothesis / probability results of the Ansari-Bradley test for SC values

h / p	Male voice			Female voice		
	Joy	**Sadness**	**Anger**	**Joy**	**Sadness**	**Anger**
Neutral [A)]	1/1.92 10^{-14}	1/2.31 10^{-14}	1/4.91 10^{-15}	1/9.3 10^{-18}	1/2.71 10^{-7}	1/1.01 10^{-12}
Joy [A)]	**0/1**	1/1.19 10^{-12}	1/2.68 10^{-12}	**0/1**	1/9.79 10^{-27}	1/4.02 10^{-12}
Sadness [A)]	–	**0/1**	1/7.11 10^{-17}	–	**0/1**	1/3.15 10^{-15}
Neutral [B)]	1/1.98 10^{-15}	1/2.04 10^{-16}	1/5.14 10^{-18}	1/1.33 10^{-28}	1/1.77 10^{-17}	1/1.32 10^{-22}
Joy [B)]	**0/1**	1/1.65 10^{-13}	1/8.04 10^{-13}	**0/1**	1/4.45 10^{-37}	**1/0.0023**
Sadness [B)]	–	**0/1**	1/7.55 10^{-23}	–	**0/1**	1/7.18 10^{-24}

[A)] Emo DB, [B)] CZ & SK stories

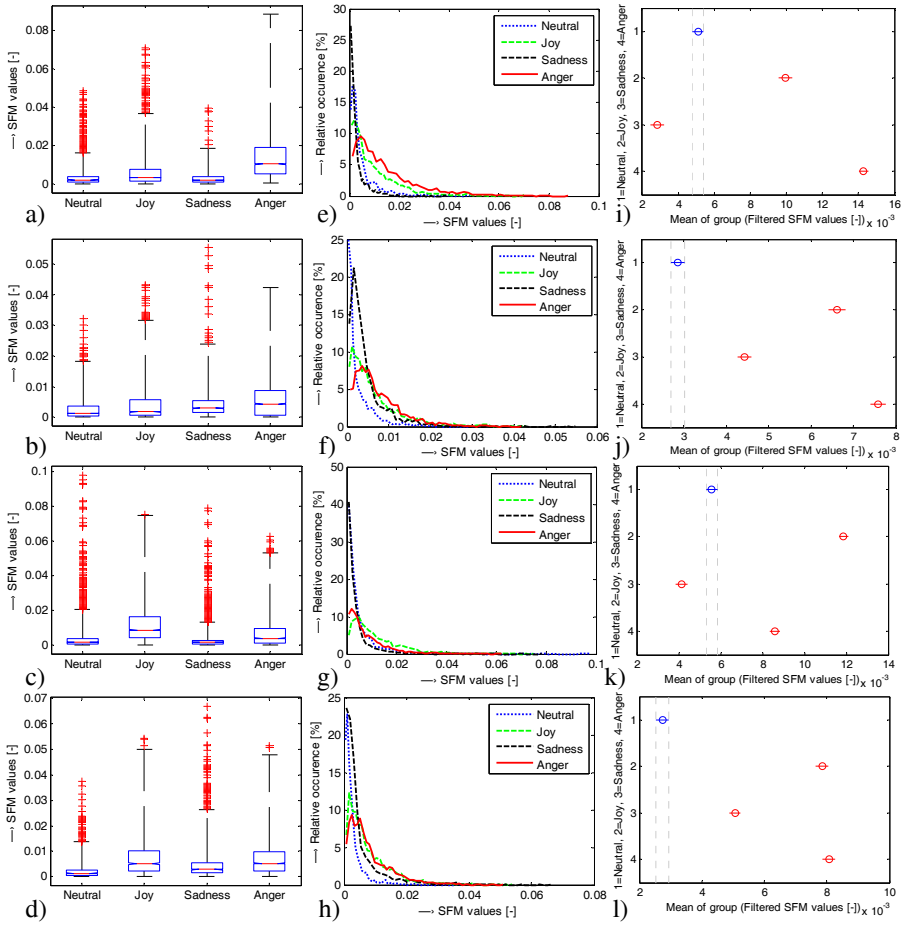


Fig. 7. Results of SFM values – boxplot of basic statistical parameters: corpus "A" male voice (a), corpus "A" female voice (b), corpus "B" male voice (c), corpus "B" female voice (d); histograms: corpus "A" male voice (e), corpus "A" female voice (f), corpus "B" male voice (g), corpus "B" female voice (h); difference between group means with the help of ANOVA statistics: corpus "A" male voice (i), corpus "A" female voice (j), corpus "B" male voice (k), corpus "B" female voice (l); determined from voiced frames only

Table 5. Null hypothesis / probability results of the Ansari-Bradley test for SFM values

h / p	Male voice			Female voice		
	Joy	**Sadness**	**Anger**	**Joy**	**Sadness**	**Anger**
Neutral [A)]	$1/1.79\ 10^{-25}$	$1/2.04\ 10^{-7}$	$1/3.69\ 10^{-15}$	$1/2.34\ 10^{-23}$	$1/1.40\ 10^{-9}$	$1/1.39\ 10^{-19}$
Joy [A)]	**0/1**	$1/1.39\ 10^{-15}$	$1/2.47\ 10^{-11}$	**0/1**	$1/2.85\ 10^{-29}$	$1/1.37\ 10^{-21}$
Sadness [A)]	–	**0/1**	$1/3.37\ 10^{-13}$	–	**0/1**	$1/4.25\ 10^{-14}$
Neutral [B)]	$1/1.52\ 10^{-12}$	$1/1.18\ 10^{-20}$	$1/1.08\ 10^{-35}$	$1/1.56\ 10^{-12}$	$1/3.28\ 10^{-35}$	$1/1.24\ 10^{-36}$
Joy [B)]	**0/1**	$1/1.71\ 10^{-28}$	$1/4.75\ 10^{-24}$	**0/1**	$1/1.32\ 10^{-15}$	**1/0.0422**
Sadness [B)]	–	**0/1**	$1/2.48\ 10^{-37}$	–	**0/1**	$1/5.15\ 10^{-24}$

[A)] Emo DB, [B)] CZ & SK stories

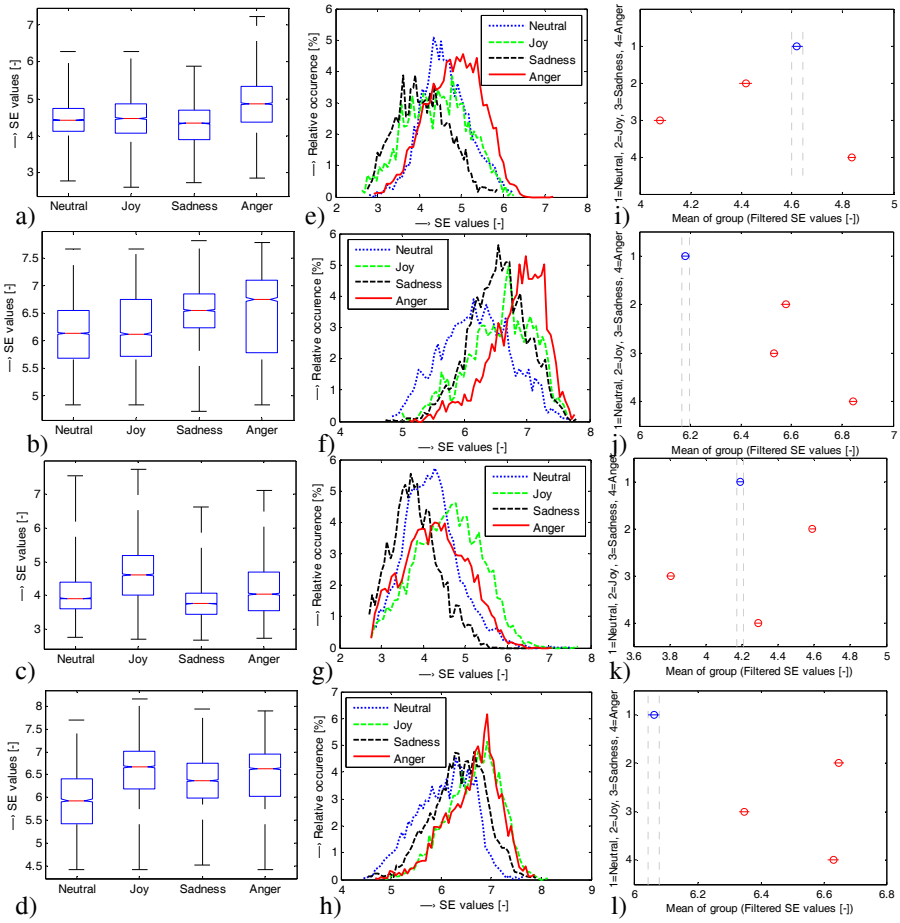


Fig. 8. Results of SFM values – boxplot of basic statistical parameters: corpus "A" male voice (a), corpus "A" female voice (b), corpus "B" male voice (c), corpus "B" female voice (d); histograms: corpus "A" male voice (e), corpus "A" female voice (f), corpus "B" male voice (g), corpus "B" female voice (h); difference between group means with the help of ANOVA statistics: corpus "A" male voice (i), corpus "A" female voice (j), corpus "B" male voice (k), corpus "B" female voice (l); determined from voiced and unvoiced frames.

Table 6. Null hypothesis / probability results of the Ansari-Bradley test for SE values

h / p	Male voice			Female voice		
	Joy	**Sadness**	**Anger**	**Joy**	**Sadness**	**Anger**
Neutral [A)]	$1/7.68\ 10^{-17}$	$1/3.74\ 10^{-35}$	$1/2.02\ 10^{-25}$	$1/2.06\ 10^{-24}$	$1/5.38\ 10^{-14}$	**1/0.0027**
Joy [A)]	**0/1**	$1/2.17\ 10^{-48}$	$1/3.28\ 10^{-36}$	**0/1**	$1/1.45\ 10^{-17}$	$1/1.08\ 10^{-16}$
Sadness [A)]	–	**0/1**	$1/4.52\ 10^{-44}$	–	**0/1**	$1/1.55\ 10^{-30}$
Neutral [B)]	$1/2.05\ 10^{-29}$	$1/9.59\ 10^{-26}$	$1/2.24\ 10^{-41}$	$1/2.44\ 10^{-15}$	$1/2.05\ 10^{-7}$	$1/7.81\ 10^{-28}$
Joy [B)]	**0/1**	**1/0.0366**	$1/2.25\ 10^{-18}$	**0/1**	$1/5.79\ 10^{-34}$	**1/0.0014**
Sadness [B)]	–	**0/1**	$1/5.56\ 10^{-19}$	–	**0/1**	$1/1.32\ 10^{-27}$

[A)] Emo DB, [B)] CZ & SK stories

Table 7. Results of basic CSF statistical analysis: mean values and standard deviations (in brackets) for male and female voices in neutral and emotional states

Emotion	Male voice			Female voice		
	SC [Hz]	SFM [x 10^{-3}]	SE [-]	SC [Hz]	SFM [x 10^{-3}]	SE [-]
Neutral A)	358.6(188.1)	3.91(0.5)	4.47(0.5)	392.4(269.8)	4.28(0.8)	4.01(0.6)
Neutral B)	321.4(163.7)	2.75(0.4)	6.12(0.5)	349.9(191.7)	2.32(0.3)	6.01(0.6)
Joy A)	466.8(262.7)	6.24(0.8)	4.44(0.6)	693.7(483.8)	7.14(0.9)	4.59(0.8)
Joy B)	403.6(236.5)	4.33(0.6)	6.22(0.6)	662.1(300.1)	7.42(0.7)	6.59(0.8)
Sadness A)	304.5(131.7)	3.09(0.3)	4.29(0.5)	290.2(188.6)	3.11(0.5)	3.76(0.5)
Sadness B)	395.7(159.7)	4.38(0.5)	6.52(0.4)	493.8(217.4)	5.18(0.7)	6.76(0.5)
Anger A)	596.5(404.7)	14.31(1.2)	4.83(0.6)	667.9(405.9)	11.92(0.9)	4.13(0.8)
Anger B)	591.9(361.9)	5.08(0.6)	6.53(0.7)	608.6(301.4)	7.16(0.7)	6.13(0.8)

A) Emo DB, B) CZ & SK stories

Table 8. Results of extended statistical analysis: skewness / kurtosis[1] parameters determined from histograms for male and female voices in neutral and emotional states

Emotion	Skewness / kurtosis for male voice			Skewness / kurtosis for female voice		
	SC	SFM	SE	SC	SFM	SE
Neutral A)	2.54 / 12.55	2.88 / 10.82	0.21 / 0.07	3.99 / 12.31	4.68 / 13.95	0.68 / 0.95
Neutral B)	2.51 / 7.56	2.56 / 8.42	0.03 / -0.56	2.91 / 7.32	3.85 / 12.58	-0.27 / -0.49
Joy A)	1.88 / 7.38	2.26 / 6.99	0.07 / -0.59	0.81 / 0.37	1.79 / 4.42	0.06 / -0.32
Joy B)	1.45 / 2.41	2.09 / 5.47	-0.41 / -0.38	0.98 / 0.72	2.02 / 5.63	-0.41 / -0.16
Sadness A)	2.81 / 8.86	3.63 / 12.36	0.38 / -0.33	2.42 / 9.89	4.41 / 17.82	0.55 / 0.21
Sadness B)	1.52 / 3.29	3.28 / 9.29	-0.24 / 0.08	1.39 / 2.01	3.32 / 14.79	-0.31 / -0.01
Anger A)	1.01 / 1.56	1.66 / 3.21	-0.15 / -0.44	0.97 / 0.46	2.24 / 6.28	0.22 / -0.58
Anger B)	0.92 / 0.83	1.84 / 4.71	-0.78 / 0.41	0.81 / 0.12	1.75 / 4.05	-0.61 / 0.21

A) Emo DB, B) CZ & SK stories

Table 9. Summary results of CSF analysis: comparison of mean value ratios between different emotional states and a neutral state for both speech databases

Mean ratio X: neutral	Joy A)	Joy B)	Sadness A)	Sadness B)	Anger A)	Anger B)
SC - male voice	1.520	1.595	1.414	1.218	2.205	2.178
SC - female voice	2.251	1.758	1.553	1.309	1.725	1.832
SFM - male voice	2.125	2.310	1.725	1.550	2.805	2.650
SFM - female voice	1.953	2.860	1.795	1.850	1.536	1.940
SE - male voice	1.093	1.064	1.090	1.056	1.036	1.107
SE - female voice	1.104	1.094	1.077	1.048	1.024	1.097

A) Emo DB, B) CZ & SK stories

[1] We use definition of kurtosis which subtracts three from the computed value, so that the normal distribution has kurtosis of zero.

4 Discussion and Conclusion

From the realized statistical analysis of the CSF values follows that our working hypothesis was confirmed. Our performed experiment confirms that for all three languages (German, Czech, and Slovak) the obtained results of the CSF values depend on a speaker but they do not depend on nationality (as well as spectral properties and prosodic parameters). The generalization of the hypothesis based on three nationalities may still be unjustified. Additionally, the corpus over which the experiment was performed is based on simulated emotions. Simulated utterances are often exaggerated and may, in many cases, differ from the utterances occurring under different real-life situations.

The ANOVA computation gives also *F* statistic and results of the hypothesis test including probability values. However, a different type of the hypothesis test was chosen in our statistical comparison of the CSF vales. Unlike the ANOVA *F* statistic, the Ansari-Bradley test compares whether two independent samples come from the same distribution against the alternative that they come from distributions having the same median and shape but different variances. From the statistical comparison realized in this way follows that there exists:

- correlation of results for male and female voices inside the currently analyzed speech corpus,
- significant differences between data groups in emotional and neutral styles.

There is also some statistical "similarity" between groups (for results obtained from the speech corpus "B"):

- "Joy" and "Anger" for female voice in the case of the SC and SFM parameters,
- "Joy" and "Sadness" for male voice, and "Joy" and "Anger" for female voice in the case of the SE parameter.

In the case of the results obtained from the SE parameter, a similarity between groups "Neutral" and "Anger" from the speech corpus "A", and between groups "Joy" and "Anger" from the speech corpus "B" for female voice, were observed (see Fig. 8j, l). Generally, it could be said that statistical distribution of groups of neutral and emotional styles taken from the speech material of the corpus "A" is better than the distribution given by the corpus "B". It would be done by the primary method of classification of sentences in emotional styles included in this corpus – in the case of the speech corpus "B", classification of emotional states was carried out manually, by subjective listening method. Also the time duration of the sentences obtained in the corpus "B" was sometimes too short for correct expression of the desired emotion.

Obtained values of the complementary spectral features together with the basic spectral properties (formant positions and their bandwidths, cepstral coefficients) and prosodic parameters (F0 contour and range, energy and duration, microintonation and jitter, etc.) can also be applied for female voice transformation and emotional speech production in a developing multi-voice TTS system [24]. As the final aim, is to use the obtained ratios of mean values to control the high frequency noise component in the speech synthesis based on statistical approach (HMM method [25]).

Acknowledgments. The work has been done in the framework of the COST 2102 Action "Cross-Modal Analysis of Verbal and Non-Verbal Communication". It has also been supported by the Grant Agency of the Slovak Academy of Sciences (VEGA 2/0090/11) and the Ministry of Education of the Slovak Republic (VEGA 1/0987/12).

References

1. Chetouani, M., Mahdhaoui, A., Ringeval, F.: Time-Scale Feature Extractions for Emotional Speech Characterization. Cognitive Computation 1, 194–201 (2009)
2. Luengo, I., Navas, E., Hernáez, I.: Feature Analysis and Evaluation for Automatic Emotion Identification in Speech. IEEE Transactions on Multimedia 12, 490–501 (2010)
3. Pao, T.-L., Chen, Y.-T., Yeh, J.-H., Liao, W.-Y.: Combining Acoustic Features for Improved Emotion Recognition in Mandarin Speech. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 279–285. Springer, Heidelberg (2005)
4. Atassi, H., Riviello, M.T., Smékal, Z., Hussain, A., Esposito, A.: Emotional Vocal Expressions Recognition Using the COST 2102 Italian Database of Emotional Speech. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) COST 2102 Int. Training School 2009. LNCS, vol. 5967, pp. 255–267. Springer, Heidelberg (2010)
5. Bozkurt, E., Erzin, E., Erdem, C.E., Erdem, A.T.: Formant Position Based Weighted Spectral Features for Emotion Recognition. Speech Communication 53, 1186–1197 (2011)
6. Iriondo, I., et al.: Automatic Refinement of an Expressive Speech Corpus Assembling Subjective Perception and Automatic Classification. Speech Communication 51, 744–758 (2009)
7. Hosseinzadeh, D., Krishnan, S.: On the Use of Complementary Spectral Features for Speaker Recognition. EURASIP Journal on Advances in Signal Processing 2008, Article ID 258184, 10 pages (2008), doi:10.1155/2008/258144
8. Berlin Database of Emotional Speech. Department of Communication Science, Institute for Speech and Communication, Technical University Berlin, `http://pascal.kgw.tu-berlin.de/emodb/` (retrieved March 13, 2006)
9. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. In: Proc. INTERSPEECH 2005, ISCA, Lisbon, Portugal, pp. 1517–1520 (2005)
10. Přibil, J., Přibilová, A.: Application of Speaking Style Conversion in the Czech and Slovak TTS System with Cepstral Description. In: Proceedings of the 14th International Conference on Systems, Signals and Image Processing (IWSSIP 2007) & 6th EURASIP Conference Focused on Speech and Image Processing, Multimedia Communications and Services (EC-SIPMCS 2007), Maribor, Slovenia, pp. 289–292 (2007)
11. Přibil, J., Přibilová, A.: Spectral Flatness Analysis for Emotional Speech Synthesis and Transformation. In: Esposito, A., Vích, R. (eds.) Cross-Modal Analysis. LNCS (LNAI), vol. 5641, pp. 106–115. Springer, Heidelberg (2009)
12. Přibil, J., Přibilová, A.: Statistical Analysis of Complementary Spectral Features of Emotional Speech in Czech and Slovak. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS (LNAI), vol. 6836, pp. 299–306. Springer, Heidelberg (2011)
13. Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions on Speech and Audio Processing 3, 72–83 (1995)
14. Hartung, J., Makambi, H.K., Arcac, D.: An Extended ANOVA F-test with Applications to the Heterogeneity Problem in Meta-Analysis. Biometrical Journal 43(2), 135–146 (2001)

15. Volaufová, J.: Statistical Methods in Biomedical Research and Measurement Science. Measurement Science Review 5(1), 1–10 (2005)
16. Vích, R.: Cepstral Speech Model, Padé Approximation, Excitation, and Gain Matching in Cepstral Speech Synthesis. In: Proceedings of the 15th Biennial EURASIP Conference Biosignal 2000, Brno, Czech Republic, pp. 77–82 (2000)
17. Li, X., Liu, H., Zheng, Y., Xu, B.: Robust Speech Endpoint Detection Based on Improved Adaptive Band-Partitioning Spectral Entropy. In: Li, K., Fei, M., Irwin, G.W., Ma, S. (eds.) LSMS 2007. LNCS, vol. 4688, pp. 36–45. Springer, Heidelberg (2007)
18. Lee, W.-S., Roh, Y.-W., Kim, D.-J., Kim, J.-H., Hong, K.-S.: Speech Emotion Recognition Using Spectral Entropy. In: Xiong, C.-H., Liu, H., Huang, Y., Xiong, Y.L. (eds.) ICIRA 2008, Part II. LNCS (LNAI), vol. 5315, pp. 45–54. Springer, Heidelberg (2008)
19. Púčik, J., Oweis, R.: CT Image Reconstruction Approaches Applied to Time-Frequency Representation of Signals. EURASIP Journal on Applied Signal Processing 2003, 422–429 (2003)
20. Kar, S., Bhagat, M., Routray, A.: EEG Signal Analysis for the Assessment and Quantification of Driver's Fatigue. Transportation Research Part F 13, 297–306 (2010)
21. Poza, J., et al.: Regional Analysis of Spontaneous MEG Rhythms in Patients with Alzheimer's Disease Using Spectral Entropy. Annals of Biomedical Engineering 36, 141–152 (2008)
22. Oppenheim, A.V., Schafer, R.W., Buck, J.R.: Discrete-Time Signal Processing, 2nd edn. Prentice-Hall (1999)
23. Boersma, P., Weenink, D.: Praat: Doing Phonetics by Computer (Version 5.2.20) [Computer Program], http://www.praat.org/ (retrieved March 25, 2011)
24. Hanzlíček, Z., Matoušek, J., Tihelka, D.: First Experiments on Text-to-Speech System Personification. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 186–193. Springer, Heidelberg (2009)
25. Hanzlíček, Z.: Czech HMM-Based Speech Synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 291–298. Springer, Heidelberg (2010)

Form-Oriented Annotation for Building a Functionally Independent Dictionary of Synthetic Movement

Izidor Mlakar[1,3], Zdravko Kačič[2,4], and Matej Rojc[2,4]

[1] Roboti c.s. d.o.o
[2] Faculty of Electrical Engineering and Computer Science, University of Maribor
[3] Tržaška cesta 23,
[4] Smetanova ulica 17,
Slovenia
izidor.mlakar@revolutionary-robotics.com,
{kacic,matej.rojc}@uni-mb.si

Abstract. Non-verbal behavior performed by embodied conversational agents still appears "wooden" and sometimes even "unnatural". Annotated corpora and high resolution annotations capturing the expressive details of movement, may improve the gradualness of synthetic behavior. This paper presents a non-functional, form-oriented annotation scheme based on informal corpora involving multi-speaker dialogues. This annotation scheme allows annotators to capture the expressive details of movement in high-resolutions. The expressive domains it captures are: spatial domain (movement-pose configuration on the level of articulators), fluidity (translations between movement-phases and phrases), temporal domain (movement variation in the form of movement phases), repetitivity (repetitive features of movement), and power (level of exposure). The presented annotation scheme can transform the encoded data into movement templates that can be directly reproduced by an embodied conversational agent.

Keywords: high-resolution movement annotation, informal corpora, non-verbal behavior reproduction, embodied conversational agents, behavior synthesis.

1 Introduction

Research into natural human-machine interaction demands a multidisciplinary effort that combines different input/output processing techniques and social sciences. Embodied conversational agents (ECAs) present artificial bodies that can, when interacting with a user, evoke some-sort of social-response. ECAs can be used within a wide range of application scenarios, e.g. intelligent interfaces, games, and educational applications. These artificial bodies can already generate both verbal and non-verbal behavior. However, there is a general agreement amongst researches that movement reproduced anon-verbal behavior still lacks naturalness [1][2]. The main reason for "unnatural" (synthetic) movement-sequences lies mainly in the fact that the motives and goals of communicative behavior are, as yet, to be completely discovered [3]. The human mechanism driving non-verbal movement (e.g. arm movement, hand gestures and facial expression) may originate from a wide variety of contexts (e.g. physiology, attitude, social relations, etc.).

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 251–265, 2012.

TV interviews and theatrical plays have shown themselves to be very usable source of conversational behavior for the analytical description of movement produced during conversation [4]. However, their degrees of spontaneity are often argued about. Therefore, more and more often researches try to rely on informal conversation in the forms of informal dialogues. There are several reasons for building corpora of spontaneous non-verbal behavior based on informal dialogs. The informal corpora may reveal additional phenomena in relations between verbal and non-verbal behavior. For instance, in informal dialogues, gestures sometimes overlap and can be perceived as a single gesture whereas, in reality, the movement segment contains several movement phrases. The fact that head and hand gestures can complement, replace or even contradict each other is even more evident in such a dialog. The phenomena observed during informal communicative dialog make information-exchange very dynamic and less predictive than in laboratory set-ups or formal TV interviews. Within the area of affective computing, this might lead the ECAs to re-produce a more credible (believable) non-verbal behavior [5].

The motivation of the presented work is towards the text-centric synthesis of more natural verbal and non-verbal conversational behavior. Text-centric applies to the fact that the behavior is synchronized with unknown input text sequences. The major disadvantage of text-centric behavior synthesis systems is the lack of the contextual information (acoustic signal information, emotion, speaker-listener relation, intent etc.). However, text-centric systems are scenario independent and can easily apply human features, such as attitude and speaker style, to any given sequence of utterances.

The verbal sequences can be synthesized as speech by using PLATTOS TTS [6], and the co-verbal movement by an embodied conversational agent EVA (ECA EVA) [7]. In order to synthesize the co-verbal movement as naturally (human-like) as possible, the aim is to transfer those dynamics and correlations, found in multi-speaker conversational dialogs, into speech-synchronized synthetic sequences of conversational behavior. The baseline of the presented research is founded on:

- a state-of-the-art TTS Engine that creates speech, prosody, and EVA-Script behavior descriptions [8],
- procedures that indicate contextual information on unknown text sequences (e.g. morphology, semantics, and syntax),
- a state-of-the-art ECA engine that can, in real time, animate EVA-Script procedural scripts.

To sum up, in order to animate (generate) co-verbal communicative behavior, it is essential to answer as to “when” and “what kind of” movements to generate. However, in order for the synthetic movement to be more believable it may even be more important to indicate the different forms of movement, and their expressive features (e.g. spatial configuration, temporal dynamics, etc.). Based on these facts, a functionally-independent annotation layer is defined that serves for coding movement details in high-resolution. The details captured by the functionally-independent layer are also used for the automatic transformation of annotated data into expressively adjustable movement templates (*gesture lexemes*). The annotation presented in this paper is based on observing and coding informal, relaxed, and multi-speaker dialogs. The

formal-model of the functionally-independent layer is form-oriented and extends the concepts presented in *form-oriented systems* (e.g. FORM [9], [10]). The functionally-dependent layer, also defined by the presented scheme, is intended to provide further insights into how the non-verbal is synchronized with verbal information. It is used to define those movement segments that serve within either the communicative or the non-communicative functions of conversation. That is, only those segments that closely relate to verbal information (e.g. *morphological relations, semantic, relations, phrases, communicative functions, iconics, dietics, etc.*) are taken into account.

This paper is structured as follows. The beginning addresses the related works. They are then followed by a brief presentation of the annotation corpora and the tools used for performing the coding of communicative behavior. Further sections then address the process of form-oriented annotation in detail. Section 5 addresses the process of automatically converting form-annotation data into Eva-Script-based behavior. Finally, the paper concludes with a section discussing the results, and a section describing our final thoughts and future work.

2 Related Work

Several annotation schemes and annotation tools have emerged in order to minimize the discrepancies in ECA's movement generation, and to provide an insight into how non-verbal behavior is structured, organized, and synchronized. Allwood in [11][12] defines two types of human mechanism for managing communication. Several functional annotation schemes have been proposed, based on Allwood's notation of conversation. . The MUMIN coding scheme [13], for instance, focuses on the annotation of three communicative functions: the feedback, turn-management, and sequencing functions. In [14], the author addresses gestures, facial expressions, and eye-gazing as non-verbal means of conveying feedback, and provides subtle cues for controlling and organizing conversations. In [15], the authors try to define those correlations between contextual factors (referent features, discourse) and gesture features, and classify them as systematic (shared among speakers), or idiosyncratic (inter-individually different). The importance of functional annotations for producing human-like non-verbal behavior using ECA's is indisputable. Such annotations offer insights into motives, and a correlation between verbal and non-verbal behavior. However, functional annotations fail to address (or only roughly) any movements' dynamically produced features, or details regarding the form of movement being produced. In order that synthetic behavior does not appear "wooden", movement's structural, power, spatial, and temporal features must also be analyzed in detail.

The annotation procedures that analyze movement at lower levels than functional annotations code the form of movement (structural, power, and spatial features) as gestural lexicons and semantic classes. The encoded data, maintained within these classifications, contains dimensions such as describing handedness, trajectory, height, distance, coarse shape and gesture space.. The temporal features of movement are coded in the form of movement phases and phrases [16, 17]. Additionally, the authors in [18] also introduce a tool for facilitating the manual annotation of gestures in video. The process of annotation they proposed, could also have a high likelihood of

successful detection should the poses be coded in an automatic way. Gestural lexicons and semantic classes can provide a gradual and human-like movement. However, the resolution at which these schemes encode movement data may be too abstract and movement-generated less reliable. In [19] a coding scheme is presented for annotating multimodal emotional behavior. The authors studied the emotional behavior in the forms of movement's expressivity (the number of repetitions, the fluidity, the strength, the speed, and the spatial expansion), the temporal features of movement's, the number of annotations in each modality, and the trajectory of the movement. Their coding scheme captures a high number of details, although only at a semi-abstract level. The gradualness and naturalness of synthetic movement performed by an ECA may as a consequence still remain a challenge.

Those annotation schemes that capture movement data at the highest-resolution are *form-oriented systems*. Form-oriented systems provide, in contrast to functional or semi-functional forms, detailed information on the spatial configuration of movement its temporal characteristics and other expressive features. The representative form-oriented systems are e.g. Ham-NoSys [20], and FORM [8]. CoGesT [21] also represents a form-oriented transcription model for hand and arm gestures, as performed during conversation. In addition to high precision (e.g. shape of the hand, size, and speed), the CoGesT scheme distinguishes between static gestures (*postures and held movements*) and dynamic gestures (*gestures with a source and trajectory).* In [10] the authors compensate for FORM's complexity by introducing a 3D pose editor integrated into ANVIL annotation tool [22]. The proposed concept allows gestures to be finely-tuned based on different-end poses and interpolated, as movement sequences, into gradual synthetic-movement.

The existing annotation approaches and schemes provide different levels for understanding the communicative human movement and its form. On the one hand, functional annotations provide detailed data on different correlations between produced non-verbal behaviors in the forms of: 1) *functions within dialog*, 2) *semantic/morphological structures* and 3) *state of the body/mind of the observant*. Functional annotations, however, only provide rough estimate about the form (*expressive features*) of movement generated. On the other hand, form-oriented annotations provide detailed data about the structure, power, and other expressive features of movement. However, their functional relations are limited to correlations with different word-phrases, at best. Semi-functional forms are the most economical and capture part of the functional, and part of the form-oriented features.

The primary goal of the presented work is the text-centric synthesis of more "natural" non-verbal human-like behavior, by using ECA EVA. The level of expressive details within abstract gestures and lexemes influences the gradualness and spontaneous appearance of synthetic movement. Merging the functional and kinematic levels of annotation as an abstraction of details may result in a faster annotation process. However, the precision of movement may be lost. This presented work differs from related research in regard to the level of detail it captures. It suggests that not only body-parts but every articulator driving the movement of a body-part should be described. Articulators are the smallest components of articulated 3D model that induce the pose overlaid by an ECA. These articulators are further grouped according to the different body-parts for which their influence may be observed (e.g. *face, hands,*

arms, head etc.). In regard to temporal precision, and in order to establish some-sort of abstract gestural dictionary (expressive re-usage), the usage of movement phases and movement phrases is implied. Movement phases group the articulators of a body-part into sequences transforming the observed body-part in parallel. Movement phrases are an extension of gesture phrases [23]. The term movement phrase simply extends the meaning of a gesture phrase. It includes any movement performed by any body-part (*not only hands*) that adds meaning to communication. Most of the related work also relies on annotating laboratory set-ups, plays, and/or formal interviews/talk shows. Such data incorporates one or two speakers, at most, with little (or no) deviations from communicative behavior "rules" (e.g. *interruptions*, *simultaneous speakers*, etc.). This work, however, targets informal and relaxed multi-speaker set-ups with high dynamics and density of spontaneous non-verbal behavior production.

3 Annotation Corpora and Annotation Tools

A multimodal corpus of spontaneous informal behavior was constructed in order to study the verbal and non-verbal relations within communicative dialog, and in order to reproduce observed non-verbal behavior using ECAs. The corpus is based on pre-transcribed TV talk-shows. The TV talk-shows are in the Slovenian language and involve a high-degree of colloquialisms.

Currently, the annotated corpora [24] contains four accurately transcribed sessions, each with durations of about 50 minutes (approximately 200 minutes of material). Each session was pre-transcribed in ELAN [25], with separate tiers for each of the participants. These tiers captured any verbal information that was produced by the speakers. An additional tier was added in order to code those utterance sequences that co-occur with movement segments. Within each session, there were five different participants; however, only two of the participants were always present throughout all four sessions, whereas the other participants were new in each talk-show. The two participants who were always present served as a "control sample" that showed whether the established relations between non-verbal and verbal behavior were consistent within different contexts, or may be a result of random chance. In each session at least 3 participants were actively contributing to the communicative dialog.

The form-oriented annotation was performed in ANVIL [22]. The primary advantage of ANVIL is its rich, hierarchically oriented tier system, in which annotators can specify the different attributes of the tiers, and the different "parent-child" relations between tiers. Since EVA-Script is also hierarchically oriented, the ANVIL tool is well suited for the purpose of form-oriented annotations and the automatic transformation of form-oriented description into non-verbal synthetic behavior. In regard to the work presented in this paper, only those dialogue acts that had relevance to the observed conversation were analyzed and coded. Each was analyzed separately, by describing the form of movement (spatial and structural features), the movement dynamics (temporal information based on the movement phases), and the key utterances that co-occured with produced movement. These features of movement were described separately for each moving body part (e.g. left-arm, left-hand, right-arm, right-hand, head, eyes, etc.).

4 Annotation: Capturing the Expressive Features of Movement

The topology and the formal model of the annotation scheme are presented in Figure 1. The scheme allows for annotating the movement lemmas (movement phases, movement phrases and movement units).

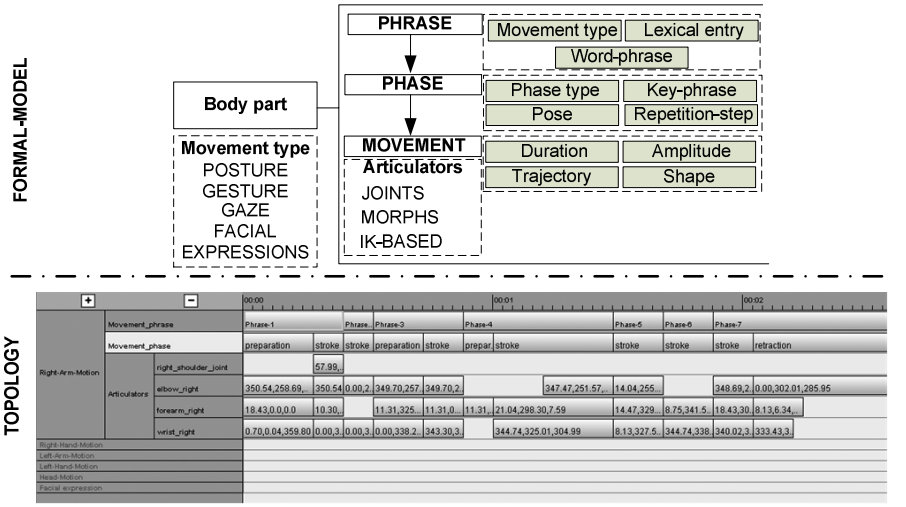


Fig. 1. The form-oriented annotation of communicative sequences in ANVIL

Topologically, the scheme shown in Figure 2, is described as a series of main tracks that hierarchically group the features of movement, based on body-parts. Therefore, body-parts are the main tracks during annotation and were defined on the skeleton configuration of the artificial ECA EVA. The body-parts are defined based on the four conversational movement types the scheme targets to analyze and code. These movement types are:

- POSTURE – describing the positions and configurations of the right and left arms
- GESTURE – describing the positions and configurations of the right and left hands
- GAZE – describing the positions of the head, and the configurations of the eyes
- FACIAL EXPRESSIONS – describing facial expressions, configurations, and emotions

The movement of the observed body-part is described in the form of *movement phrase, movement phase*, and the *articulators* propagating the observed movement. Each of these descriptors is defined as a separate track within the scheme's topology and also defines several attributes that capture the expressive details of movement. The *movement phrase* (Figure 2) (similarly to the gesture phrase [16, 17]) describes the full span of phases (from preparation to retraction). Each movement phrase contains a mandatory stroke and optional preparation, hold, and retraction phases. Movement phrase, therefore, joins sequential movement phases into continuous movement and also joins sequential *movement phases* into lexical groups that can be

reused during the annotation and animation processes. The *word phrase* attribute indicates those propagated utterance-sequences that co-occur during movement propagation. Currently, it is only used for easier identification and movement labeling. However, in the future, an utterance-sequence may indicate the correlation between a movement phrase and the general text (e.g. the relation between what the ECA expresses and the text sequences it speaks).The *movement types*, used to identify movement in its abstract form are mostly derived at based on McNeill's works on hand-gestures [16].

The *movement phase* describes the temporal and repetitive features and spatial configurations of those articulators influencing the observed movement. A movement phase is defined by its pose and phase type. The pose identifies an abstract description (e.g. *closed-fist*, *hand-wave*) of the movement being propagated. The phase type identifies the phase (preparation, stroke, retraction and hold) in which the movement is located at a given time. Each movement phase at its borders (beginning/end), therefore, defines an end-pose of the observed body-part. The proposed annotation scheme is, therefore, also pose-oriented and only the spatial configurations of key-poses are coded. A key-pose is further described by the features of movement being propagated. The movement is defined in the form of articulators (elements propagating the observed movement); their spatial configuration (shape and trajectory), the amplitude of configuration and the time period required for propagation between end-poses. The movement phases also define the animation's key-frame interpolation type (*Fluidity expressive dimension* [26]) to be used whilst animated by ECA EVA.

As described in [7], EVA-Framework supports three types of articulators, *the joint-based articulators*, *the morphed-shape-based articulators* and *the inverse-kinematics-based (IK) articulators.* Joint-based articulators are mostly used when generating head, eye, and arm and hand synthetic movement (e.g. posture, gaze and hand-gestures). The spatial configuration of such articulators is modeled by modifying their Heading, Pitch, and Roll values. The morphed-shape-based articulators are mostly used whilst generating facial expressions, emotions and other facial and eye region configurations (e.g. eye blinking, raising brows, etc.). The inverse-kinematics-based (IK) articulators require the least coding time. They offer a limited set of full body-part spatial configurations. For instance, by positioning one IK articulator within the 3D space, a proper HPR configuration of all of the right-arm's articulators can be achieved. However, the (IK) articulators are based on inverse kinematic rules, and therefore offer only a limited set of body postures.

4.1 Movement Phases and Phrases

Most of the information, carried by co-verbal movement, is presented through stroke and pre/post-stroke holds. However, form-oriented annotations can also tell us a lot about the preparation and retraction phases. The events prior and post to stroke are also important when recreating annotated movement. Namely, these events carry part of the dynamical features and also information on the best way to transit from one movement phrase to another. For instance, the temporal variations of preparation and the movements' retraction phases may be important for their re-creation based on the

general input text. The border phases may also indicate the notion of mental processes such as thinking, consideration, hesitation etc. If these two phases are also described, the ECA may better simulate: (1) word-based indicators for the start and end of the movement, (2) transitions between different movement phrases.
In order to code movement phases based on movement observation, the definitions provided by [17] were modified as follows:

- *movement phrase*: segment of movement that maintains the general trajectory. It ends in stroke, hold, or retraction phases.
- *stroke*: phase of movement, where the dynamics and shape are manifested by the greatest clarity of movement (the part of the movement is manifested with the most energy).
- *preparation*: phase of movement that leads-up to the stroke (initiates stroke). It is usually identified by slower, less-energetic movement.
- *recovery(retraction)*: phase of movement that transfers the gesture into a relaxed or withdrawn state. The state is maintained for longer periods, and can only be followed by another *stroke*/*preparation* phase.
- *post-stroke hold (hold)*: arrives at the end of the stroke, as a nucleus of the gesture phrase. It can be followed by stroke or retraction phases.

As shown in the formal-model of the scheme (Figure 1) the relations between the movement phrase, movement phase, and articulators are of a hierarchical nature. In general, the movement phrases are parents of the movement phases. Each movement phrase contains at least one stroke-phase and no more than five optional phases. Optional retraction and preparation phases may be observed on the borders of the movement phrase. In many cases, the end of preparation and the start of the beginning phase may be unclear. The movement phase should then be coded as a stroke.

Annotating Movement Phases
A movement phase defines a key-pose of the observed segment (start and end pose) and its temporal attributes; the temporal structure, and the dynamics of the movement. The *stroke phase* is, in general, defined based on the significance of movement (if a movement is more energetic then its attribution to verbal information is more significant). The pre/*post-stroke-hold* phase is defined on those segments where the pose pre/after the stroke phase remains relatively static (retains its stroke-based end-pose). The existence of hold phase suggests that the stroke and the co-expressive speech express an idea created in advance. The retraction phase labels those movement segments that drive the observed body part into a relaxed (neutral) state. It also indicates those movement segments that drive the observed body part into a static configuration that is maintained for longer periods of time (e.g. the pose of the body part is static until the end of the dialogue act). Finally, the preparation phase denotes those segments that drive the observed body part into a stroke phase. Within this phase the movement is prepared for, withheld if need be until the co-expressive speech is ready (can indicate hesitation).

Whilst observing actual movement sequences, it can be seen that the preparation phase commonly unites with the stroke phase; the borders between these phases can't be clearly defined by movement observation (e.g. fast arm movement, slow head movement, etc.). In such cases it is assumed that the movement phrase has no preparation phase and that the preparation phase will, if necessary, be appended at the phrase classification level (based on similar phrases and their word relation). Figure 2 presents how the movement phases are identified on some actual movement sequences (based on the observation of movement).

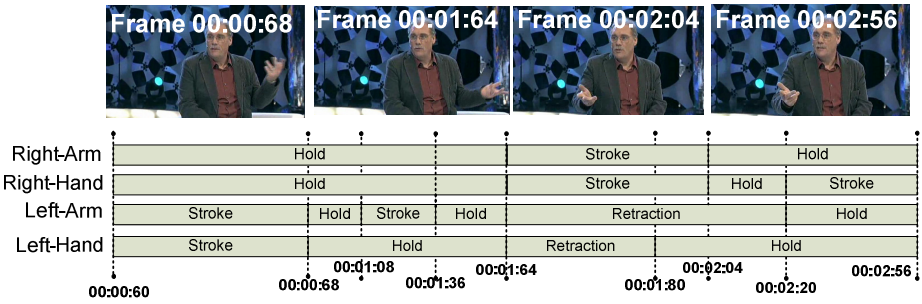


Fig. 2. Identification and coding of movement phases

Figure 2 presents a composition of continuous movement phases within a movement segment, annotated as defined by the described formal model. An end-pose defines the form of the observed body-part at the end, or at the beginning of the *movement phase*. The temporal values in Figure 2 are relative to the beginning of the observed movement segments. Figure 2 further demonstrates the importance of the proposed body-part configuration and the process of coding the movement phases. It can be seen that arm and hand movements are *mostly* correlated. E.g. at the beginning of the *retraction phase* the hand correlates with the retraction of the arm. However, when looking at the intervals between *00:01:64* and *00:02:20*, it can be observed that the retraction phases of the left arm and hand do not match. The left hand retracts faster than left arm. Another interesting fact is shown at the interval from *00:00:60* to *00:01:64*. It shows how the left hand retains its pose in a *post-stroke-hold*; however, the left arm changes its spatial configuration in the form of a *hold-stroke-hold* sequence.

When movement is reproduced, based on the general input text, combining the functional level of annotation and the movement phases/phrases may produce those word triggers relevant to spatial configuration and the dynamical features of movement. A detailed analysis of movement can further elaborate the exact temporal relations between utterances (e.g. *words*, *word phrases)* and co-verbal behavior. For instance, if similar movement phrases are grouped into lexical classes (e.g. gesture/movement units,), it can be extrapolated as to which movement type is more biased towards utterances, communicative functions, or even to sequencing of movement phases, etc.

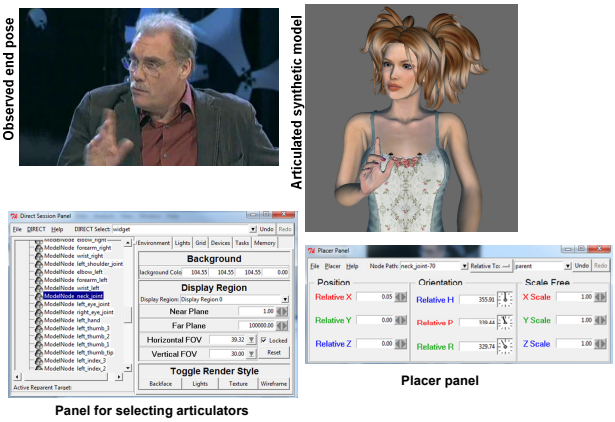


Fig. 3. Coding the end-pose by visual comparison

4.2 Annotating Spatial Configuration

Figure 3 presents the *observed end-pose* (top-left-hand-side), the approximation of the pose on the articulated model (top-right-hand-side), and the EVA-Framework's interface for the manual formation of end-poses (bottom-right and bottom-left -hand-sides). Figure 3, therefore, represents how the form of the shape is encoded within the context of ECA EVA. Each body part may contain several articulators. However, the spatial features are only described for those articulators that propagate movement during the observed movement phase. The different end-poses observed individually for each body part. Each defined body part has a pre-defined set of articulators that control its spatial configuration and movement. For instance, an arm's movement is controlled by a *shoulder-joint, elbow-joint, forearm-joint* and *wrist-joint*, whereas the hand is controlled by *outer-* and *inner-palm-joints*, and three joints for each finger. The configuration of articulators mostly depends on the articulated model of the synthetic agent. Therefore, the concept of the proposed annotation scheme can easily be adapted to any type of articulated model. The process of coding the spatial features of movement based on direct visual comparison is as follows: Firstly, the annotator approximates the pose of the synthetic model to the overlay of the observed end-pose. The annotator selects the articulators from the list of articulators provided by the *panel for* selecting *articulators* (Figure 3 left-hand side) and modifies its spatial features until the two poses match. When the modeling of the end-pose is finished, the annotator has to transcode the appropriate HPR values into the annotation scheme. Optionally the annotators can, as described by the formal model of the scheme, also specify in an abstract form, as to the "shape"/"pose"/''lexical entry'' of similar previously-coded movement. In addition, If the head-movement is being described, the annotators can also (optionally) describe the direction of gaze. If gazing is not provided, either by articulators or gaze attribute, it will be automatically adjusted by the EVA Framework.

Currently, the proposed coding process is still quite time-consuming. It may take several hours of adjustment before all the key poses of the movement segment are properly described. However, the detail captured and the precision of the reproduced movement already outweigh the cost of coding. In addition, any movement phrase or movement segment may be expressively reused, as described in [7]. This means that the articulated synthetic agent will be able to reproduce movement that can vary in *temporal*, *power,* or even *repetitive dimensions* of expressivity. Also the dynamics of the reused movement phrase are automatically adopted, based on annotated movement phases. Each movement phrase can also be used in several combinations, in order to form new movement segments. In order to minimize the effort of the coding process, end poses can also be defined in external 3D modeling software, such as: Maya3D[1], Daz3D[2], Blender[3], etc.

Encoding Facial Expressions

Facial expressions defined within the EVA-Framework are based on facial action points (FAPs) and predefined facial expressions. Although the annotators could code each FAP as a separate articulator, an annotator agreement had been reached that simplifies the annotation and makes it less bias. The facial expressions are described either as emotions (e.g. sad, happy, angry, etc.) or as finite, limited expression-sets performed around the mouth (e.g. smile, opened mouth, puckered lips), cheek (e.g. puffed cheeks), and eyebrows (raised or lowered) regions. The models for defining and describing expressions are based on the MMI facial expressions database [27] and the level of exposure ranges from 1-10.

5 Transforming Annotated Data into Movement Models

EVA script and EVA framework [7] support the concepts of expressive movement and expressive movement models. The EVA script's movement descriptions are hierarchically oriented, and provide several expressive attributes compatible with the previously discussed form-oriented annotation scheme. The transformation process, from scheme to movement model, can form movement phases, movement units, and also complete movement segments. All are described in EVA Script and can be used directly in the reproduction of movement on an embodied conversational agent. Figure 4 demonstrates how annotated data for the right-arm is automatically transformed into Eva-Script's description.

The process of transforming the annotation of body parts firstly identifies the annotated movement phases. An XML template in EVA Script (*a movement template*) is formed for each identified movement phrase.

The *movement phases* define the overall temporal features of the phrase and the type of key-frame interpolation. The preparation movement phase denotes *"easeIn"* interpolation, the retraction phase denotes *"easeOut"* interpolation, the hold phase

[1] Maya 3D – `http://usa.autodesk.com/maya/`
[2] Daz3D – `http://www.daz3d.com/i/products/daz_studio?`
[3] Blender – `http://www.blender.org/`

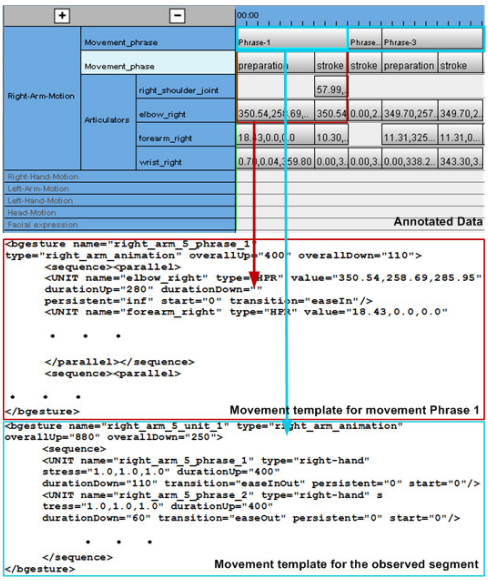


Fig. 4. Transformation of annotation into movement templates for EVA-framework

denotes "*easeOut"*, and the *stroke phase,* no animation interpolation. The hold phase also defines movement that is maintained over a certain time. It is reflected as the *persistency* expressive feature of movement. Movement phases are transformed into EVA-Script's "<sequence><parallel>" blocks, and movement templates for poses.

The tracks for articulators then define the spatial configuration of the propagated movement. These tracks are processed individually and inserted within the corresponding movement template blocks (EVA-Script's "<sequence><parallel>" blocks) as EVA Script's UNIT tags. In addition to the spatial expressive dimension, the articulators can, within the temporal borders of movement phase, also define their own "local" temporal features. For instance, if the modeling of an articulator is delayed, a start attribute of the UNIT tag is set. The annotated articulators and their attributes are processed individually for each phase block.

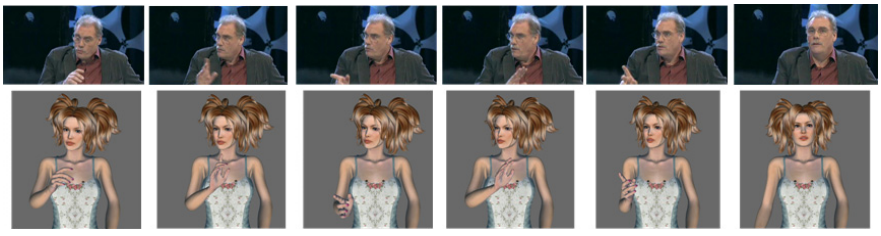


Fig. 5. Reproduced speaker imitation

Figure 5 presents how the form-oriented movement annotation is reproduced as synthetic movement, as imitated by ECA EVA. The presented image sequence contains several sequential frames, as collected from the original video sequence (upper sequence), and the animated synthetic video sequence (lower sequence).

6 Results

Initially, 37 minutes of informal conversation have already been annotated. This annotated material involved four speakers, 3 male and 1 female, actively participating in the dialog. 534 single and multi-speaker segments were identified that showed significant hand, arm, or head movement attributed to co-expressive speech. The annotators were asked to disregard any segments that could present a random movement (e.g. random head movement, sitting down, changing position due to physical discomfort, etc.); any movement that cannot clearly be related to the spoken content or communicative functions. Each movement segment was further segmented based the on observation of the movement. Namely, some of those segments identified as relevant for communication, were broken into shorter samples. The borders of the movement segment were finalized as timestamps, where the movement returned to a rest pose or the pose was maintained for longer periods. The movement segments were also evaluated by visually comparing the original sequences and those synthetically produced based on the annotated values. Whilst synthesizing the annotated data (Figure 5) the form and dynamics closely matched to as performed by human speakers. The proposed annotation scheme and reproduced results were quite encouraging.

7 Conclusion

This paper described, in detail an annotation scheme that could be used to build a high resolution, functionally-independent movement dictionary for the reproduction of communicative behavior. We have discussed the annotation scheme, and the reproduction of annotated data using synthetic agent EVA.

The topology of the presented annotation scheme extends the general hand-gesture oriented topologies in several ways. Firstly it adopts the notion that any type of body movement can be regarded as a carrier of meaning. It is designed in such a way that enables posture, gesture, gaze and facial expressions to be described within a single session and under a shared time-line. The shared timeline enables the annotators to establish relations between different movement types, especially between arm-posture and hand gestures. Secondly, the annotation scheme also defines those word-phrases and key-phrases to be captured during the same annotation session.

Currently, annotated movement still represents a small part of the dictionary the ECA should use. We therefore intend to annotate the entire multimodal corpora (200 minutes), and if necessary, additional video samples of informal dialog will be added. Furthermore, as indicated in [28], form-oriented annotations are quite time consuming (e.g. 20 hours of coding per 1 minute of video for FORM). Our approach is estimated to take 5-7 hours per 1 minute of video. In contrast to [28] (1 hour of coding per 1

minute), this value is still relatively high. Therefore, we are planning to upgrade the system, in order that it would be able to automatically approximate the spatial attributes of control units based on pose recognition and tracking techniques (e.g. [29]). The annotators would then only have to code movement phases and phrases, and check the correctness of the recognized movement. This should reduce the annotation process time and would be much more time-efficient.

Acknowledgements. Operation part financed by the European Union, European Social Fund.

References

1. Foster, M.E., Oberlander, J.: User preferences can drive facial expressions: evaluating an embodied conversational agent in a recommender dialogue system. J. of User Modeling and User-Adapted Interaction 20(4), 341–381 (2010)
2. Novielli, N., de Rosis, F., Mazzotta, I.: User attitude towards an embodied conversational agent: Effects of the interaction mode. J. of Pragmatics 42(9), 2385–2397 (2010)
3. Read, S.J., Talevich, J., Walsh, D.A., Chopra, G., Iyer, R.: A Comprehensive Taxonomy of Human Motives: A Principled Basis for the Motives of Intelligent Agents. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 35–41. Springer, Heidelberg (2010)
4. Martin, J.C., Caridakis, G., Devillers, L., Karpouzis, K., Abrilian, S.: Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviors in TV Interviews. In: Maglogiannis, I., Karpouzis, K., Bramer, M. (eds.) Artificial Intelligence Applications and Innovations. IFIP, vol. 204, pp. 369–377. Springer, Boston (2006)
5. Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., Pantic, M.: A Multimodal Database for Mimicry Analysis. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 367–376. Springer, Heidelberg (2011)
6. Rojc, M., Kačič, Z.: Time and space-efficient architecture for a corpus-based text-to-speech. Speech Communication 49(3), 230–249 (2007)
7. Mlakar, I., Rojc, M.: Towards ECA's Animation of Expressive Complex Behaviour. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 185–198. Springer, Heidelberg (2011)
8. Rojc, M., Mlakar, I.: Multilingual and Multimodal Corpus-Based Text-to-Speech System - PLATTOS. Speech Technologies/Book 2 (2011)
9. Martell, C.: Form: An Extensible, Kinematically-Based Gesture Annotation Scheme. In: Advances in Natural Multimodal Dialogue Systems, vol. 30, pp. 79–95 (2005)
10. Nguyen, Q., Kipp, M.: Annotation of Human Gesture using 3D Skeleton Controls. In: Proc. of the Seventh International Conference on Language Resources and Evaluation, LREC 2010 (2010)
11. Allwood, J.: Dialog Coding – Function and Grammar. Gothenburg Papers in Theoretical Linguisics 85 (2010)
12. Allwood, J., Ahlsén, E., Lund, J., Sundqvist, J.: Multimodality in own communication management. Current. Trends in Research on Spoken Language in the Nordic Countries 2, 10–19
13. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. J. of Language Resources and Evaluation 41(3), 273–287 (2007)

14. Jokinen, K.: Gaze and Gesture Activity in Communication. In: Stephanidis, C. (ed.) UAHCI 2009. LNCS, vol. 5615, pp. 537–546. Springer, Heidelberg (2009)
15. Bergmann, K., Kopp, S.: Systematicity and Idiosyncrasy in Iconic Gesture Use: Empirical Analysis and Computational Modeling. In: Kopp, S., Wachsmuth, I. (eds.) GW 2009. LNCS, vol. 5934, pp. 182–194. Springer, Heidelberg (2010)
16. McNeill, D.: Gesture and Thought. University of Chicago Press (2005)
17. Kendon, A.: Gesture: Visible action as utterance. Cambridge University Press (2004)
18. Zhang, J.R., Kuangye, G., Herwana, C., Kender, J.R.: Annotation and taxonomy of gestures in lecture videos. In: Proc. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1–8 (2010)
19. Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C.: Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 405–417. Springer, Heidelberg (2005)
20. Prillwitz, S., Leven, R., Zienert, H., Hanke, T., Henning, J.: HamNoSys Version 2.0. Hamburg Notation System for Sign Languages (1989)
21. Trippel, T., Gibbon, D., Thies, A., Milde, J.T., Looks, K., Hell, B., Gut, U.: CoGesT: A Formal Transcription System for Conversational Gesture. In: Proc. of LREC 2004 (2004)
22. Kipp, M.: Anvil - A Generic Annotation Tool for Multimodal Dialogue. In: Proc. of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367–1370 (2001)
23. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press (1992)
24. Verdonik, D., Zwitter-Vitez, A., Romih, M., Krek, S.: Konkordančnik za govorni korpus GOS = Concordancer for the speech corpus GOS. In: Proc. of the 13th International Multiconference Information Society - IS 2010, vol. C, pp. 12–15 (2010)
25. Sloetjes, H., Russel, A., Klassmann, A.: ELAN: a free and open-source multimedia annotation tool. In: Proc. of INTERSPEECH 2007, pp. 4015–4016 (2007)
26. Hartmann, B., Mancini, M., Pelachaud, C.: Towards affective agent action: Modelling expressive ECA gestures. In: Proc. of International Conference on Intelligent User Interfaces (2005)
27. Pantic, M., Valstar, M.F., Rademaker, R., Maat, L.: Web-Based Database for Facial Expression Analysis. In: Proc. of Multimedia 2005, pp. 317–321 (2005)
28. Kipp, M., Neff, M., Albrecht, I.: An annotation scheme for conversational gestures: how to economically capture timing and form. J. of Language Resources and Evaluation 41(3), 325–339 (2007)
29. Peng, G., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: Proc. of Computer Vision 2009, pp. 1381–1388 (2010)

A Cortical Approach Based on Cascaded Bidirectional Hidden Markov Models

Ronald Römer

TU- Cottbus, Chair of Communication Engineering,
Konrad Wachsmann Allee 1, 03046 Cottbus, Germany
ronald.roemer@tu-cottbus.de

Abstract. Research in the field of neural processing proposes a bidirectional computation scheme among the hierarchical organized levels of the brain. This scheme is called cortical algorithm and can be realized using Cascaded Bidirectional Hidden Markov Models (CBHMMs). In this paper CBHMMs are investigated in the light of analysis-synthesis systems. Such systems are important elements of Cognitive Dynamic Systems and Cognitive User Interfaces. Some of the most salient properties of Cognitive Systems are their abilities to support inference and reasoning, planning under uncertainty and adaptation to changing environmental conditions. That is, beside the bidirectional computation scheme among the hierarchical organized levels, CBHMMs need to support logical operations like inference and reasoning. To integrate this new aspect to the analysis-synthesis framework we pick up an old suggestion from D.M. MacKay from the late 1960s. D.M. MacKay suggested to supplement Shannon's measure of selective information content by a descriptive information content. Descriptive information in turn is composed of structural and metric information and considers the logical aspect of information.

Keywords: Cognitive Dynamic Systems, Analysis-Synthesis Systems and Cortical Algorithm.

1 Introduction

The human brain is undoubtedly the most powerful cognitive dynamic system in our biological world. As a salient property the distributed feedback may be emphasized as a fundamental principle of biology. This principle is embodied in the cybernetic cycle of interactions of the brain with its environment. The cybernetic circle describes the perception of the environment along the sensory hierarchy and the action on the environment along the motor hierarchy. Both hierarchical organized structures are indicated by a bidirectional flow of information and level specific working memories [1]. If inspiring ideas arising by looking to the human brain and they could be summarized in a powerful computational model as well, then a new generation of engineering systems enabled with cognition can be expected [2]. First conceptual attempts in this direction are made in [3] for instance.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 266–272, 2012.

Otherwise, research in the field of neural processing suggests increasing evidence that the neocortex of the brain does not consist of a collection of specialized and dedicated cortical architectures, but instead possesses a fairly uniform, hierarchically organized structure. This uniformity implies that the same general computational processes are performed across the entire neocortex, even though different regions are known to play different functional roles [4]. Recently, CBHMM and multirate-CBHMM structures were introduced, which represent the working memories for each hierarchical level [5]. Based on these structures the so called cortical algorithm has been formulated. In this algorithm both the analysis- and the synthesis path are using predictive information, that stem from neighboring hierarchic levels. Hence, a bidirectional flow of information takes place at the same time and is fused in each level according to the Bayesian principle mentioned above. A consistent justification of the simultaneous bidirectional flow of information is given by the usage of inner models of communication participants [6].

This paper is organized as follows. Initially, in section 2 we are focusing on the properties and requirements to Cognitive Systems and motivate the usage of CBHMMs in Probabilistic Hierarchical Bidirectional Analysis Synthesis Systems. Subsequently, the Bayesian Inference Mechanism is introduced and it's implication for CBHMMs as a realization of an analysis-synthesis system is commented. Finally, the information theoretical interpretation and a conclusion is given.

2 Cognitive Dynamic Systems and Cognitive User Interfaces

Recently S. Haykin proposed to combine model based signal processing techniques with new ideas from neurosciences. The resulting systems are called Cognitive Dynamic Systems [2]. In such systems the transmitter and the receiver are linked by the environment and uses feedback to optimize communication. Feedback is used to describe the human cognition by the cybernetic cycle which has close connections to the system theory or to the theory of closed loop control systems.

Moreover, the adaptation to changed environment conditions and the handling of uncertainties need to be considered by cognitive systems. To realize such a cognitive system, the regarding analysis-synthesis system should have the following major features: a hierarchical bidirectional structure according to the biological model, the motor and the sensory hierarchy process probability distributions, all hierarchical levels use only one unique algorithm and finally the algorithm follows the principle of sequential Bayesian Filtering or Kalman Filtering respectively.

A further approach with a cognitive background was proposed by Steve Young. He states, that future human-machine interfaces should exhibit the following key characteristics [7]: the ability to support reasoning and inference, the user interface must be capable to interpret inputs robustly to resolve ambiguities and the system should be able to plan under uncertainty.

3 The Cortical Algorithm Based on CBHMMs

To explain the basics of the cortical algorithm, firstly the relation between HMMs and Kalman Filtering or Bayesian Filtering is explained. An HMM may be separated into the static part, which is used to compute the observation probability and the dynamic part which is modeled by state transitions. Further, it's important to note that an HMM process probability distributions like a Kalman Filter. That is, firstly a prediction based on the dynamic part is computed, moreover the likelihood of the observation is computed. Finally, the posteriori distribution is derived by the combination of both parts followed by a normalization. This relationship is mathematically described by the generalized Kalman equation:

$$p(z_k|\boldsymbol{y}_{0:k}) = \frac{p(\boldsymbol{y}_k|z_k) \cdot p(z_k|\boldsymbol{y}_{0:k-1})}{p(\boldsymbol{y}_k)}. \tag{1}$$

By the introduction of further additional HMM-layers there is the opportunity to involve context from higher levels (analysis) or to involve feedback from lower levels (synthesis). The coupling of states of different levels can be understood as a spatial transition from one quality of states to another quality of states. To expand the generalized Kalman equation to an hierarchical system, an additional third term needs to be fused according to the Bayesian principle to get the a-posteriori distribution of states.

CBHMM structures are extensively investigated in [5]. Hence, at this point only the fusion equation is emphasized at this place

$$\gamma_j^d(k) \propto \left[\sum_l \beta_l^{d-1}(k) \cdot b_{j,l}^d\right] \cdot \left[\sum_i \gamma_i^d(k-1) \cdot a_{i,j}^d\right] \cdot \left[\sum_r \alpha_r^{d+1}(k) \cdot b_{j,r}^{d+1}\right]. \tag{2}$$

In this equation three components are combined to get the a-posteriori state distribution at every level d and at each instant k. The temporal prediction based on $\gamma(k-1)$ from level d, the bottom-up component: $\alpha(k)$ from level $d+1$ and the additional third term, the top-down component: $\beta(k)$ from level $d-1$ (see figure 1). In the next section the distinct meanings of the third term for the analysis- and synthesis stage are investigated.

4 Bayesian Inference Mechanism

By the introduction of the HMM-technology some important aspects of signal modeling have been considered. On the hand the static and dynamic aspects at state level are modeled by the generalized Kalman equation. On the other hand the transition from the subsymbolic level to symbolic levels is modeled by soft quantization and by the transition from classes to states. In [6] these details were mathematically explained. The most interesting part from the point of view of CBHMMs is the transition from classes to states as represented by the following equation

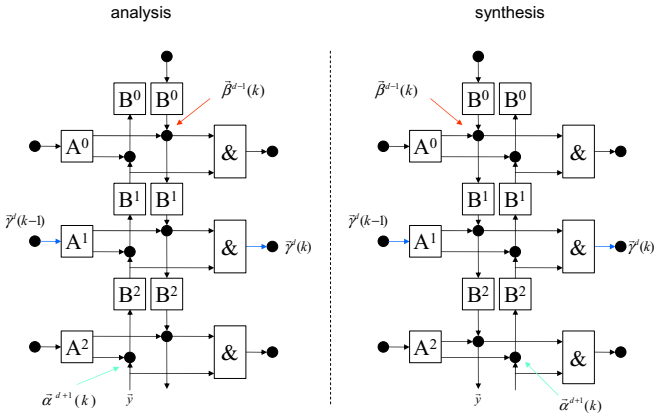


Fig. 1. Analysis-synthesis system based on a CBHMM structure: transitions within the hierarchical levels are described by A-Matrices. Transitions between hierarchical levels are modeled by B-Matrices. The fusion of the bidirectional flow of information is indicated by the AND-symbol.

$$b_z(\boldsymbol{y}) = \sum_R c_{z,r} \cdot N(\boldsymbol{y}|\boldsymbol{\mu_{r,z}}, \boldsymbol{\Sigma_{r,z}}), \tag{3}$$

where the number of classes or prototypes are denoted with R and the states by z. The term $b_z(\boldsymbol{y})$ describes the emission distribution for the observation vector $\boldsymbol{y}$ given the state z, weighted by a sum of Gaussian mixtures. Gaussian mixtures are using classes ω_r which are determined by a mean vector $\boldsymbol{\mu}$ and by the respective covariance matrix $\boldsymbol{\Sigma}$. The equation above is understood as soft interpretation and is closely related to the Bayesian inference mechanism. The classical notation of inference rules and probabilistic rules respectively are given by

$$\frac{A, A \overset{rule}{\longrightarrow} B}{B} \quad \text{and} \quad \frac{A, A \overset{P(B|A)}{\longrightarrow} B}{B}. \tag{4}$$

That is, if a rule is applied to the fact A then the result B can be concluded. The Bayesian inference mechanism uses a slightly different notation: each rule gets a weight or a probability respectively. The undirected joined probability of two random variables may be computed in two directions, in forward direction $p(A, B) = p(B|A) \cdot (A)$ and in backward direction $p(A, B) = p(A|B) \cdot (B)$. If we have A_k facts and B_i results, then we can compute the regarding conditional probabilities

$$p(B_i) = \sum_K p(B_i|A_k) \cdot p(A_k), \quad \text{and} \quad p(A_k) = \sum_I p(A_k|B_i) \cdot p(B_i). \tag{5}$$

Both equations have the same mathematic structure as in (3). Hence, the weights $c_{r,z} = p(s_z|\omega_r)$ in (3) may be interpreted in the same way as the probabilistic rule $\omega_r \xrightarrow{c_{r,z}} s_z$ according to (4). That is, the mapping process from the symbol quality 'class' to the symbol quality 'state' by probabilistic rules corresponds to a soft interpretation process. In opposite to the prediction part in (2) - where the symbol quality 'state' is not changed by horizontal transitions - the vertical transitions change the symbol quality. Furthermore it is worth to remark, that in the reverse horizontal direction other weights are used as in the bottom-up direction. Hence, to utilize the same weights for the reverse direction the Bayesian relation may be applied again. This scheme is called Bayesian inference mechanism:

$$p(A_j|B_i) = \frac{p(B_i|A_j) \cdot p(A_j)}{\sum_K p(B_i|A_k) \cdot p(A_k)}. \tag{6}$$

According to the terminology of reasoning and inference the top-down process corresponds to a process which is called 'deduction' whereat the bottom-up process is denoted as 'abduction'. At this point a new aspect of information - the logical aspect - appears, but this aspect is not captured by Shannon's information theory so far. Shannon's basic idea is sustained by the fact, that transmitter and receiver just need to know symbols from a common shared finite alphabet. Hence, no information for the reconstruction of symbols is needed, but rather selective information is required. Unfortunately, the question how the selection process is controlled, was not answered by Shannon's information theory.

5 Information Theoretical Interpretation

D.M. MacKay has tackled the control problem by the question how the receiver may recover the meaning of a message [8]. To answer this question, he firstly adopted the concept of 'states of conditional readiness'. This concept is based on the observation, that it may happen in the course of a communication process, that a message seems not to change the behavior of the communication participant. What has been affected by the participant's understanding of the message is not necessarily an observable response but rather what the communication participant would be ready to do if relevant circumstances arose in the close future. With other words, the communication participant has just changed his inner state.

That is, it is not the behavior of the participant, but rather the state of conditional readiness which may change in the course of the communication process. Hence, D.M. MacKay concluded that the meaning of a message can be defined as its selective function on the range of the receiver's 'states of conditional readiness' for goal directed behavior. A change in meaning implies a different selection from the range of states of readiness. A meaningless message is one that makes no selection from that range. By this way of thinking a conceptual bridge between

mechanism and meaning is proposed. Further, it offers a criterion of meaningfulness and meaninglessness. For the transmitter, the intended meaning of the message is the selective function he wants to perform on the receiver's range of states of readiness. This is distinct and may be different from the cognitive meaning of the receiver; and both of these may differ from the conventional meaning, which is the selective function calculated for a 'standard receiver'. Both ideas, the descriptive information and the selective function on the range of 'states of conditional readiness' help to interpret CBHMMs and emerge from the fusion equation. The latter will be explained subsequently.

To extend the information theory, D.M. MacKay introduced the descriptive information which can be further decomposed to metrical and structural information. The metrical information corresponds to the number of atomic facts or the 'weights of evidence', this kind of information has at first no relation to the logical dimension. It can represent any kind of logical entities. In contrast to the metrical information a relationship to the logical dimension may be established by the structural information. That is, using structural information one can infer to the logical degree of freedom and to get measurable observations descriptive information is needed. This point of view can be generalized to probability distributions and then applied to CBHMMs.

Metrical information is achieved by the process of soft vector-quantization. The number of atomic facts corresponds to the number of classes or codebook prototypes. Based on the observation of a feature vector each class is then occupied by it's 'weight of evidence'. If structural information is added by using probabilistic rules, then the emission probability for each state can be estimated. The occupation of the classes may have completely different meanings, the meaning of this occupation arise only by the computation of the weighted sum (see equation 3). Hence, this process is called soft-interpretation. In CBHMMs the soft interpretation takes place in all levels along the bottom-up direction. The probability distribution of one level provides metrical information for the next higher level. By applying the probabilistic rules, the probability distribution of the next level may be interpreted. That is, an inference mechanism (abduction) comes to existence, whereat more and more abstracted facts are propagated to higher levels, so that an analysis-synthesis system can work on different time scales. In the reverse direction an inference mechanism is existing as well (deduction), but in the top down direction the abstracted facts are more and more expanded to get finer details.

As mentioned above, in biological models and of course in CBHMMs the analysis- and the synthesis stage as well is characterized by an bidirectional flow information at the same time. Therefore it is necessary to combine both streams of information at the their respective levels by the fusion equation. Afterwards the posterior-distributions of states is available and the selective function on the 'level specific range of states' may follow by the transmitter or receiver. The selective function may be realized by the well known Viterbi-decoding algorithm or by A*-search.

6 Conclusion

The fusion equation (2) contains two independent components: the a-priori selective uncertainty - $\gamma(k-1)$ from level d - and the descriptive information from the abduction process - $\alpha(k)$ from level $d+1$ - or from the deduction process - $\beta(k)$ from level $d-1$ - respectively. In each hierarchical level both components are fused to get the a-posteriori selective uncertainty. That is, the a-priori uncertainty - delivered by the predictions in each level - is reduced by the descriptive information that stem from the abduction and deduction processes along the hierarchy. Hence, one can conclude that the descriptive information controls the selective function in the transmitter as well as in the receiver. This finding is compatible to the idea of D.M. MacKay, that the meaning of a message is determined by the selective function of the communication participants whereat the selective function in turn is controlled by the descriptive information.

References

1. Fuster, J.M.: Cortex and Mind. Unifying Cognition. Oxford Press (2003)
2. Haykin, S.: Cognitive Dynamic Systems. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, vol. 4, pp. 1369–1372 (2007)
3. Mumford, D., Lee, T.S.: Hierarchical bayesian inference in the visual cortex. Journal of the Optical Society of America 20(7) (2003)
4. Mountcastle, V.: An Organizing Principle for Cerebral Function: The Unit Model and Distributed Systems. In: Edelman, G.M., Mountcastle, V.B. (eds.) The Mindful Brain (1978)
5. Roemer, R., Herbig, T.: Konzeptionelle Beschreibung des Corticalen Algorithmus und seine Verwendung in der automatischen Sprachverarbeitung. In: 20. Konferenz Elektronische Sprachsignalverarbeitung, ESSV 2009, pp. 33–40. TUD Press, Dresden (2009) (in German)
6. Roemer, R.: Beschreibung von Analyse-Synthese-Systemen unter Verwendung von CBHMM's. In: 22. Konferenz Elektronische Sprachsignalverarbeitung, ESSV 2011, pp. 67–76. TUD Press, Aachen (2011) (in German)
7. Young, S.: Cognitive User Interface. IEEE Signal Processing Magazine, ICASSP-2007 27(3), 128–140 (2010)
8. MacKay, D.M.: Information, Mechanism and Meaning. MIT Press (1969)

Modeling Users' Mood State to Improve Human-Machine-Interaction

Ingo Siegert, R. Böck, and Andreas Wendemuth

Otto von Guericke University Magdeburg, Germany
ingo.siegert@ovgu.de

Abstract. The detection of user emotions plays an important role in Human-Machine-Interaction. By considering emotions, applications such as monitoring agents or digital companions are able to adapt their reaction towards users' needs and claims. Besides emotions, personality and moods are eminent as well. Standard emotion recognizers do not consider them adequately and therefore neglect a crucial part of user modeling.

The challenge is to gather reliable predictions about the actual mood of the user and, beyond that, represent changes in users' mood during interaction. In this paper we present a model that incorporates both the tracking of mood changes based on recognized emotions and different personality traits. Furthermore we present a first evaluation on realistic data.

Keywords: Emotion, Mood, Personality, Simulation of Affect, Human-Machine-Interaction.

1 Introduction/Motivation

In the future, technical systems will act as a companion [5], being able to adapt themselves to individual skills and preferences and recognize the emotional state of a user. For this, the research community has been focusing on emotion recognition and provides remarkable results. Today, the community is able to recognize emotional episodes from speech, mimics, and biosignals with reliable confidence [6,2,17]. This enables us to build emotion-aware computers, that recognize emotions and react with rule-based dialogue strategies. But to develop "affective computers", something more is necessary as Picard [15] points out. Rules alone are not sufficient to understand or predict human behavior and intelligence. Besides observations and/or expert given rules, a description about the inner mental state of the user is necessary as well.

A first step towards this could be the mood modeling to incorporate the user's emotional development during system interaction. This makes it possible to evolve a user model and gives the opportunity to predict the continuous development of the interaction. The presented technique allows us to incorporate the temporal emotional development into a mood model, predicting and identifying changes in the emotional trend. In [1] and [9] similar approaches modeling virtual characters able to react in an emotionally natural way are presented.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 273–279, 2012.

As stated in [3], emotions reflect short-term affects, usually bound to a specific event, action, or object. Hence, an observed emotion reflects a distinct user assessment, that is related to a specific occurred experience. Therefore, the system cannot conclude a rising danger of dialog abortion only from one negative emotion observation. However, ongoing negative observations could indicate this.

In contrast to emotions, moods reflect medium-term affects, generally not related to a concrete event [14]. They last longer and are more stable affective states influencing the user's cognitive functions directly. Personality, in contrast, reflects a long-term affect and individual differences in mental characteristics. A common representation scheme is the Five Factor Model of Personality (see [11] and [13]), using five traits to specify a general behavior. A predicted user mood in combination with personality can be used to draw a conclusion about the course of the conversation and the risk of dialog-abortion [7].

Therefore, we combine recognized emotions and given personality traits to model the user's mood to finally get a prediction about the progress of Human-Machine-Interaction (HCI). The main influences on our considerations can be found in [4] and [8]. They are focused on the development of human-like behavior of avatars, whereas we want to model user dispositions within HCI.

The remainder of the paper is structured as follows: In Sect. 2 we present our mood model in detail, introduce mood transitions, and integrate the user's personality. Subsequently, in Sect. 3 we will present an evaluation of our model. In Sect. 4 we briefly conclude our ideas and then give an outlook on our next research activities.

2 Modeling the Mood

As stated in the Sect.1, we rely on the user mood to get an indicator of the disposition within actual HCI. In this section we go deeper into the technical aspects of our research and focus on some phenomena. The principal questions we want to answer are:

- How to describe the mood of a user.
- How to model mood transitions caused by emotions.
- How to incorporate personality traits.
- How to constitute moods as a medium term affect.

2.1 Describing the Mood of a User

As stated in [13] the mood can be illustrated by a dimensional model using pleasure, arousal, and dominance (PAD) as dimensions. Given that, emotions can be displayed within this space, too. In our implementation the PAD mood space uses the axes +P for pleasant and -P for unpleasant, +A and -A for aroused, and unaroused and +D and -D for dominant and submissive.

The emotions of a user can already be detected with acceptable results by body reactions, vocal response patterns, and facial expressions (c.f. [16]). However, it is not possible, or not known, how to deduce information about the user's

mood directly. Hence, the mood has to be derived implicitly from observed emotions and can thus only be regarded as an assumption. Another possibility is to use questionnaires, which is however, not suitable for automatic systems in HCI.

To illustrate the impact of recognized emotions on mood, we modeled the observed emotions e at time t as forces F_t^e, which perform a distinct shifting ΔL_M of the mood M within the PAD-space. The absolute value of F_t^e is equivalent to the distance from the point of origin to the observed emotion (the norm of the emotion vector) multiplied with a user independent factor κ_0 in the interval of $(0, 1]$. To decrease the effect of a single emotion force, this effect is weakened by a distinct modifiable damping term D. The following terms model the indirect impact of emotions to mood changes:

$$F_t^e = \kappa_0 \cdot e \tag{1}$$

$$\Delta L_M = \frac{F_t^e}{D_{t-1}} \tag{2}$$

$$M_t = M_{t-1} + \Delta L_M \tag{3}$$

$$D_t = f(F_t^e, D_{t-1}, \mu_1) \tag{4}$$

The mood update works as follows, (see Fig. 1): The actual emotion force F_t^e is derived from the observed emotion e and user independent factor κ_0. It is used to update the mood M by calculating ΔL_M utilizing the previous damping D_{t-1}. The actual damping D_t is updated as well, by using the actual emotion force F_t^e and the previous damping D_{t-1} (see Sect. 2.2).

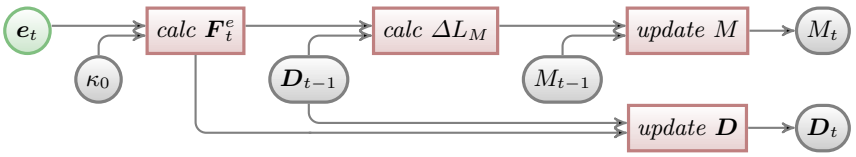


Fig. 1. Scheme of our mood model. Grey circles are inner model values, green circle is an observed emotion, red boxes are calculations.

2.2 Modeling Mood Transitions Caused by Emotions

As stated before, a damping term was introduced to decrease the impact of the emotional force. Furthermore, we used directional damping terms D^p, D^a, and D^d to investigate the effects for each dimension separately. Therefore, we also examined the emotion force independently for all three parameters: F^p, F^a and F^d. So, we are capable to model both, mixed emotions and multiple simultaneously recognized emotions. This allows us to deal either with several expressed

emotions at the same time, which can be be split into distinct emotions, or handle emotions that are represented just along one axis.

Additionally, the damping term depends on sign. In the direction of emotion force ($\rightrightarrows$) the damping is decreasing, in the opposite direction ($\rightleftarrows$) the damping is increasing. Thereby, we model the effect that a cumulation of similar emotions tends to pull the mood into their direction within the PAD-space. Simultaneously, the chance that the mood moves towards the direction of the actual emotion decreases. Hence, we use two damping values for every dimension of the PAD-space, one in direction of emotion and one against it. The value of the damping term is changed by every mood update. Thus it comprises the development of observed emotions. The appearance of the function was chosen in such a way, that it reflects the following effects: by defining D_{max} the strength of the damping can be controlled and by defining μ_1 the gradient can be controlled. The damping depends only on the value of the previous damping D_{t-1} and the actual emotional force F_t^e.

This modeling technique is turn-based, as mentioned in Sect. 2.1. It follows directly from the definition of moods as a mid-term affect, which means that moods tend to change slowly. We also implemented a so called "fall-back-emotion". Every time-step where no emotion is observed, a default emotion is applied to the model. This emotion represents the user's initial mood state and thereby shifts the mood step-wise back to its initial state.

2.3 Incorporating Personality

As mentioned in Sect. 1, personality traits play an important role in HCI as well. To incorporate this phenomenon in our model, we chose two main ideas that have strongest impact on mood perception, as stated in [12] and [8]: (i) The determining of an initial mood and (ii) the translation of an observed emotion into an emotional force.

Different users can have individual attitudes towards technical systems caused in parts by their different personalities, which can be represented by the Five Factor Model. In [13] a mapping of these personality traits into the PAD-space is presented. We use this representation and place the initial mood into the region or quadrant that is represented by the relationship analog to [12].

The second phenomenon is the possibility to adjust the observed emotions from different users. We noticed, similar to [10], that observed emotions from different users with the same intensity can be felt by the users themselves in a totally different manner. Dependent on such personality, the way emotions are presented can vary. For this, a translation of observed emotion into their internal representation is needed. Focusing on the intensity we are using a factor to determine the difference between observation and internal feeling of the user's emotion. We utilizing the personality trait "extraversion" that influences our adjustment factor κ_η.

Figure 1 incorporates these issues: The observed emotions e_t and derived emotion forces F_t^e are represented as a vector containing PAD-dimensions. Whereas the damping D_t is represented as a matrix containing sign dependent PAD-dimensions.

3 Experimental Evaluation

For the evaluation of our model we used data from the experiment described in [17]. This corpus, called EmoRec-Woz I, was generated during a Wizard-of-Oz experiment where the users had to play games of concentration (Memory). It contains audio, video, and bio-physiological data. Each experiment was divided into two rounds with several ESs. The experiment was designed in such a way that through feedback, wizard responses, and game difficulty different emotional states were induced. In result, this corpus contains 10 sessions for both rounds of about 30 minutes length. During the experiment the user was supposed to pass several octands in the PAD-space. The sequence of the experimental sequences (ESs) and expected PAD-positions is shown in Table 1. For each sequence several triggers were used, to induce emotions like difficulty of used card set, positive or negative feedback.

Table 1. Sequence of ES and expected PAD-positions

ES	Intro	1	2	3	4	5	6
PAD	all	+++	+-+	+-+	-+-	-+-	+-+
mood prediction	–	↗	↗	↗	→	↓	↑

We then used the triggers, given by wizard, as input for an appraisal process to "form" emotions, see [16]. We did not get the real emotion felt by the user, but an idea, what and when he could have felt emotions. Thereby it is possible for us to form a realistic development. In ES1, ES2, ES3, and ES6 mostly positive emotions, in ES4 and ES5 mostly negative emotions were induced. Investigations with emotion recognizers using prosodic, facial and bio-physiological features [17] and the comparison to the experimental design support this emotional course. As we could not give a realistic examination for all emotions, we concentrated on pleasant and unpleasant. Using this data as input to our presented mood model, we were able to show, that our model follows the prediction for pleasure of given ESs in the experiment, see Fig. 2. In the beginning of ES1 the mood rests in its initial position and it takes some time-steps, until the mood starts to shift towards the positive region. In ES2, and ES3 the mood continues to rise. During ES4, when inducing more negative emotion spikes, the mood slowly moves backwards. At the end of the also negative ES5 the mood is negative, which continues at the beginning of ES6. Furthermore, during the course of ES6, where many positive emotions are induced, it is possible to change the mood again towards positive pleasure.

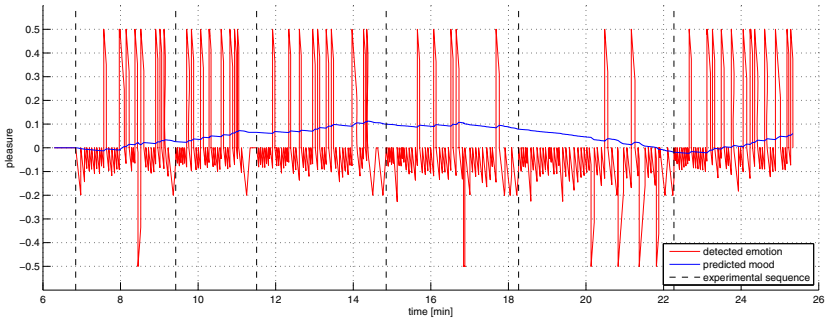


Fig. 2. Evaluation of proposed mood model using realistic data of one experimental session

4 Conclusion and Outlook

In this paper we presented a new modeling technique that allows predictions about a user's mood. We used definitions from [13] and considered moods as a result of damped influences by emotional forces. We also provided an evaluation of our model on realistic data. The next step would be to use real data to show that the proposed model is able to give a valid prediction for the user's mood.

Our technique is able to derive moods from emotions and place both into the PAD-space. Emotions act as forces onto the mood. By a changeable damping, we modeled the effect of a mid-term mood that is affected by emotions only indirectly. It is also possible to include personality in our model as well. On the one hand, by adjusting the force of an emotion dependent of the users "extraversion" value. On the other hand, by starting with different initial moods.
In contrast to other solutions like [4] and [8], which are intended to model embodied affective characters, we model the users' mood seen from the system's view. Moreover we do not need to specify appraisal rules beforehand, as we can use the output from emotion recognizers in addition or as a substitution. Furthermore, we also include a relaxation phase. If no emotion is actually recognized, a "fall-back-emotion" is used to model the mid-term effect tending the mood to fall back into its initial state.
Applying this model to technical systems, predictions of intentions or recognition of emotional behavior on a mid-term level can be accomplished. For this, the user's moods, instead of the user's emotions which could be influenced more by surrounding factors, have to be inferred.

Acknowledgement. This research was supported by grants from the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. André, E., Klesen, M., Gebhard, P., Allen, S., Rist, T.: Integrating Models of Personality and Emotions into Lifelike Characters. In: Paiva, A.C.R. (ed.) IWAI 1999. LNCS, vol. 1814, pp. 150–165. Springer, Heidelberg (2000)
2. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N.: Whodunnit – Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech. Computer Speech and Language 25(1), 4–28 (2011)
3. Becker, P.: Structural and relational analyses of emotions and personality traits. Zeitschrift für Differentielle und Diagnostische Psychologie 22(3), 155–172 (2001)
4. Becker-Asano, C.: WASABI: Affect Simulation for Agents with Believable Interactivity. Ph.D. thesis, Universität Bielefeld (2008)
5. Biundo, S., Wendemuth, A.: Von kognitiven technischen Systemen zu Companion-Systemen. KI - Künstliche Intelligenz 24, 335–339 (2010)
6. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proc. of the 6th International Conference on Multimodal Interfaces, New York, pp. 205–211 (2004)
7. Davidson, R.J.: On emotion, mood, and related affective constructs. In: Ekman, P. (ed.) The Nature of Emotion: Fundamental Questions. Oxford University Press (1994)
8. Gebhard, P.: ALMA A Layered Model of Affect. In: 4th International Joint Conference of Autonomous Agents & Multi-Agent Systems, pp. 29–36 (2005)
9. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A conversational agent as museum guide: design and evaluation of a real-world application, pp. 329–343. Springer (2005)
10. Larsen, R.J., Fredrickson, B.L.: Measurement issues in emotion research. In: Kahneman, D., Diener, E., Schwarz, N. (eds.) Well-being: Foundations of Hedonic Psychology, pp. 40–60. Russell Sage Foundation (1999)
11. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. Journal of Personality 60(2), 175–215 (1992)
12. Mehrabian, A.: Analysis of the Big-five Personality Factors in Terms of the PAD Temperament Model. Australian Journal of Psychology 48(2), 86–92 (1996)
13. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology 14(4), 261–292 (1996)
14. Morris, W.N.: Mood: the frame of mind. Springer (1989)
15. Picard, R.W.: Affective Computing. MIT Press, Cambridge (2000)
16. Scherer, K.R.: Appraisal considered as a process of multilevel sequential checking. In: Scherer, K., Schorr, A., Johnstone, T. (eds.) Appraisal Processes in Emotion: Theory, Methods, Research, pp. 92–120. Oxford University Press, New York (2001)
17. Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C., Schwenker, F.: Multimodal Emotion Classification in Naturalistic User Behavior. In: Jacko, J.A. (ed.) HCII 2011, Part III. LNCS, vol. 6763, pp. 603–611. Springer, Heidelberg (2011)

Pitch Synchronous Transform Warping in Voice Conversion

Robert Vích and Martin Vondra

Institute of Photonics and Electronics, Academy of Sciences of the Czech Republic,
Chaberska 57, CZ 18251 Prague 8, Czech Republic
{vich,vondra}@ufe.cz

Abstract. In this paper a new voice conversion algorithm is presented, which transforms the utterance of a source speaker into the utterance of a target speaker. The voice conversion approach is based on pitch synchronous speech analysis, Discrete Cosine Transform (DCT), nonlinear spectral warping with spectrum interpolation and pitch synchronous speech synthesis with overlapping using the speech production model. The DCT speech model contains also information about the phase properties of the modeled speech frame, but is, in contrary to a model based e.g. on the discrete Fourier transform, a real model and can be efficiently used for speech coding and voice conversion. The resulting finite impulse response of the converted DCT speech model is obtained by the inverse DCT and it is of the mixed phase type. The proposed voice conversion procedure results in speech with high naturalness.

Keywords: speech analysis, frequency transformation, voice conversion, speech synthesis.

1 Introduction

Voice conversion is a speech signal processing tool for achieving a change of the voice identity. The aim of the voice conversion is to convert the timbre and the suprasegmental parameters of the source speaker's voice into another one's that can be perceived either as that of a known target speaker or as of a new speaker. Voice conversion is mainly intended for Text-to-Speech systems (TTS) to provide several different types of output voices without the creation of a new speech inventory for a new speaker, which is a highly complex and time consuming procedure. From this perspective, it would be better if we had the possibility to directly influence and modify the individual characteristics of the speech model inclusive the style of speaking. In this paper a voice conversion approach based on Discrete Cosine Transform (DCT) is presented.

Already in 1995, voice conversion has been the main topic of the special issue of Speech Communication [1]. An extensive bibliography on voice transformation may be found in the dissertation of Kain [2]. Voice conversion has been also for several years in the centre of interest in the Institute of Photonics and Electronics, AS CR. Different approaches have been studied, e.g. a nonlinear frequency scale mapping

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 280–289, 2012.

combined with spline interpolation and implemented using the harmonic speech model [3], cepstral speech synthesis [4] and last but not least using PSOLA and resampling [5]. In [6] a new computationally effective and efficient voice conversion procedure based on the application of the cepstral vocoder was presented. In [7] the speech generation model based on the complex cepstrum was described. This modeling approach takes into account not only the speech magnitude spectrum, but also the phase properties of the speech signal. For that reason the speech signal is approximated with high accuracy. The decomposition of the speech signal using the complex cepstrum into the maximum- and minimum-phase components has been used in [8] for the modification of the glottal flow signal with the aim to obtain synthetic emotional speech.

The basic speech production model is based on the source-filter theory (Fig. 1). In the simplest case the excitation is represented by Dirac unit impulses with the period equal to the fundamental period of speech for voiced sounds and by white noise for unvoiced speech. The vocal tract model is represented by a time varying digital filter, which performs the convolution of the excitation with its impulse response. The vocal tract model can be based on linear prediction [9], on approximation of the inverse cepstral transformation [10], using the impulse response obtained by complex cepstrum deconvolution [7], etc.

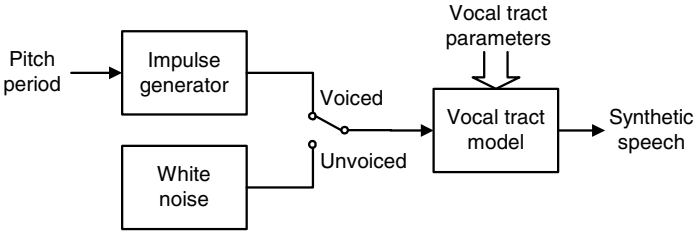


Fig. 1. Parametric speech production model

The vocal tract model is a time varying digital filter usually based on:

- Linear prediction. In this case it is of the infinite impulse response (IIR) type with *minimum-phase* and with poles only.
- The real cepstrum. Then it is also of *the minimum-phase* IIR type, but with poles and zeros. The real cepstrum speech modeling approximates only the speech magnitude spectrum.
- The complex cepstrum, which results in a finite impulse response (FIR) filter with *mixed phase*. The complex cepstrum speech model approximates not only the spectrum magnitude, but also the phase of the speech spectrum.

Linear prediction and cepstrum based speech modeling are competitive approaches for speech compression in the sense that both reduce the redundancy of the signal. The cepstrum based speech modeling is in principle a form of transform speech coding based on repeated Discrete Fourier Transform (DFT). The complex cepstrum speech synthesis is of the mixed phase type and for that reason the speech can be generated with higher

naturalness. It allows also the decomposition of the speech signal into the minimum- and maximum-phase components [8]. The minimum-phase part can be considered as the vocal tract impulse response and the maximum-phase part is first of all given by the open phase of the glottis. Shortly, the maximum-phase part of the speech signal can be considered as the excitation, i.e. as the glottal signal.

If we withdraw the decomposition of the speech signal into its components, we can use another transform coding, the DCT [11]. It is only slightly suboptimal in performance compared with the Karhunen-Loeve transform and it is very fast. The DCT has been used in adaptive transform speech coding e.g. by Zelinski and Noll in 1977 [12] and Tribolet and Crochiere in 1979 [13].

2 DCT Speech Coding

Transform speech coding based on DCT is a method for efficient speech signal transmission. It offers the possibility to reduce the redundancy of the speech signal using adaptive quantization of the DCT model coefficients. The basic transform speech coding procedure is shown in Fig. 2.

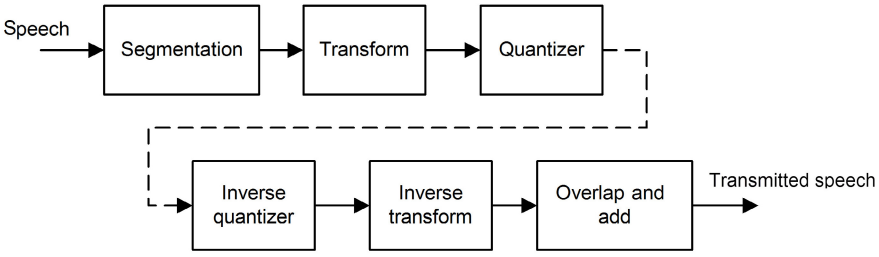


Fig. 2. Speech transform coding

The transform operation can be realized using DFT or DCT. Both approaches are similar, because they transform the signal into frequency components and have *unique inverses*. The DCT is related to the DFT and can be defined using DFT. But the DCT model is *real* and more suitable for compression and coding than the DFT model, which is *complex*. The DCT reduces the information in the frequency domain into fewer components and the real DCT components are easier to use and quantize.

In [7] we used for speech transform coding the complex cepstrum and for efficient speech synthesis pitch synchronous speech segmentation with overlapping based on pitch pulses localization for the estimation of glottal closure instants. The modeled speech frame consisted of two pitch periods, where the central pitch pulse was in the middle of the frame. Frame shift was set to one pitch period. Then a weighting window was applied on the speech frame. The type of the window has a great impact on the spectrum of the speech frame. The widely used Hamming window is in this case not suitable, because it does not suppress, in the case of voiced speech, the

periodicity of the frame completely. A better choice is the Hann or Blackman window. The pitch synchronous segmentation and the used window cause that the magnitude spectrum is very smooth – the periodicity of the voiced excitation is totally destroyed and the magnitude spectrum approximates the speech spectrum envelope.

There are several types of DCT, see [11]. Further we shall use the unitary DCT-2, which is more suitable for signal compression. The DCT-2 is defined by

$$\begin{aligned} S_k &= w_n \sum_{n=0}^{M-1} s_n \cos\frac{\pi k(2n+1)}{2M}, \qquad k = 0,1,...,M-1, \\ w_n &= \sqrt{1/M}, \qquad n = 0, \\ w_n &= \sqrt{2/M}, \qquad n > 0. \end{aligned} \tag{1}$$

The sequence $\{s_n\}$ is the windowed speech frame of the length N appended with zeros. The DCT is computed for M points, $M \geq N$. The scaling factor $\{w_n\}$ produces frequency components in a range of values similar to that of the signal components. The inverse DCT-2 is given by

$$s_n = \sum_{n=0}^{M-1} w_n S_k \cos\frac{\pi k(2n+1)}{2M}, \qquad n = 0,1,...,M-1. \tag{2}$$

As an example we use the stationary part of the vowel *a* with the fundamental frequency $F_0 = 118$Hz sampled with the sampling frequency $F_s = 8$kHz. The signal is pitch synchronously weighted using the Hann window, centered on the central glottal closure instant, with the frame length N of two fundamental periods, where N is the integer value $N = \lfloor 2F_s / F_0 \rfloor = 135$, see Fig. 3. The dimension of the DCT is $M = 512$.

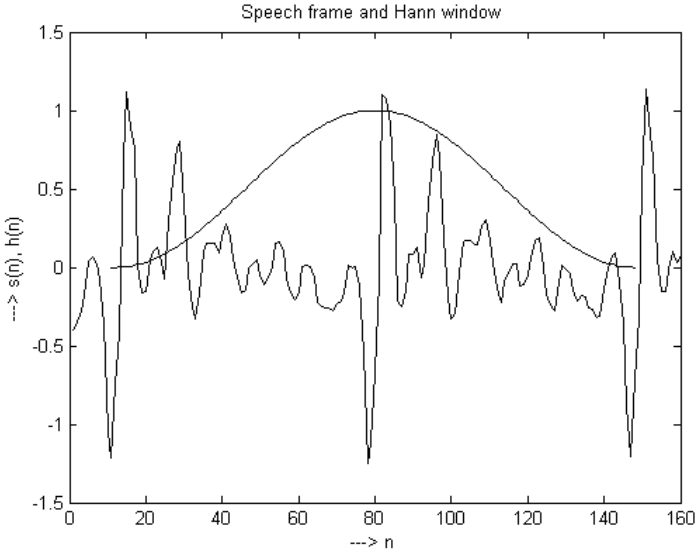


Fig. 3. Speech signal and the Hann window centered on the glottal closure instant

In Fig. 4 the DCT spectrum and DFT magnitude spectrum are shown. The DFT magnitude spectrum, which can be considered as the approximation of the vocal tract magnitude frequency response, is the envelope of the DCT spectrum. The phases in the DCT components are only 0 or π, but the phase of the speech frame spectrum is encoded into the waves of DCT spectrum.

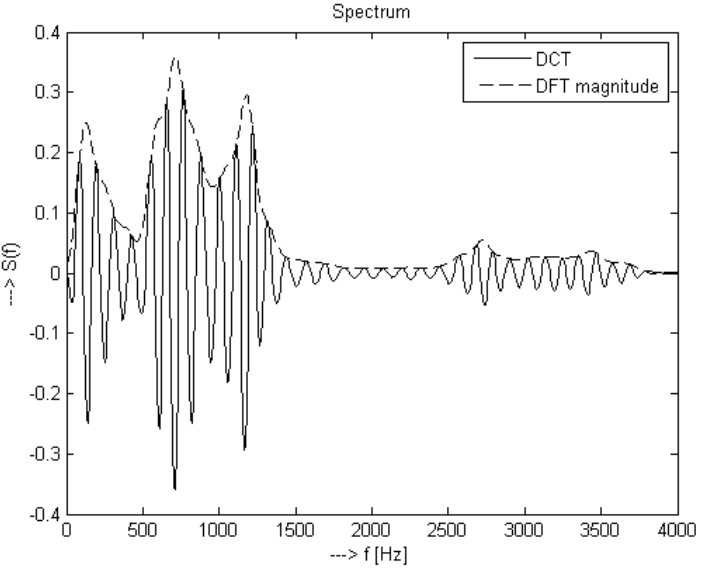


Fig. 4. DCT spectrum and the DFT magnitude spectrum

3 Speech Spectrum Modification

For voice conversion we transform the speech spectrum using the frequency warping [6]. The warping let be given by the requirement

$$S_T(f) = S_S(F), \quad 0 \le f \le F_s/2, \quad 0 \le F \le F_s/2, \tag{3}$$

where $S_T(f)$ and $S_S(F)$ are the short time spectra of the *target* speaker and the *source* speaker respectively. The variables f and F are the corresponding frequency variables. In DCT spectrum analysis the source spectrum $S_S(F)$ is given equidistantly for

$$F_k = k\frac{F_s}{2M}, \quad k = 0,1,2,...,(M-1). \tag{4}$$

F_s is the sampling frequency and M is the dimension of the DCT. The transformation of the source speaker spectrum $S_S(F)$ into the target speaker spectrum $S_T(f)$ is provided by mapping of the variable F into the variable f, i.e.

$$f = Q\ (F). \tag{5}$$

This function may be given numerically by a table or analytically and generally it is nonlinear. That means that even if F is equidistantly sampled, f is not equidistantly sampled. Therefore, the transformed S_T (f) must be equidistantly interpolated for further application at the points

$$f_k = k\frac{F_s}{2M}, \quad k = 0,1,2,...,(M-1). \tag{6}$$

In general, frequency transformations in digital filter synthesis can be obtained using a transformation similar to the bilinear transformation. In our case of voice conversion we need not transform the transfer function of the digital vocal tract model, we may apply only the frequency scale warping corresponding to the lowpass-to-lowpass transformation defined by

$$Z = \frac{z-\alpha}{1-\alpha z}, \tag{7}$$

where $Z = e^{j\Omega}$, $z = e^{j\omega}$, $\Omega = 2\pi F / F_s$, $\omega = 2\pi f / F_s$ and α is the transformation parameter.

The function (7) transforms a lowpass filter with the cutoff frequency $F = F_c$, the source filter, into a new lowpass filter, the target filter, with the cutoff frequency $f = f_c$ If we set e.g. the cutoff frequency F_c equal to the 1st formant frequency of the source speaker, i.e. $F_c = F_1$ and the cutoff frequency f_c equal to the 1st formant frequency of the target speaker, i.e. $f_c = f_1$, then the transformation parameter α is given by

$$\alpha = \frac{\sin(\pi(F_1 - f_1)/F_s)}{\sin(\pi(F_1 + f_1)/F_s)}. \tag{8}$$

The frequency warping function $f = Q(F)$ is then

$$f = F + \frac{F_s}{\pi}\arctan\frac{\alpha\sin(2\pi F/F_s)}{1-\alpha\cos(2\pi F/F_s)}. \tag{9}$$

This lowpass-to-lowpass nonlinear frequency warping function for F_1 = 1000Hz, and f_1 = 1200Hz, corresponding to $\alpha = 0.1032$, is depicted in Fig. 5.

Using the frequency warping function (9) the converted DCT model can be constructed. The original and the warped and interpolated DCT spectra are shown in Fig. 6.

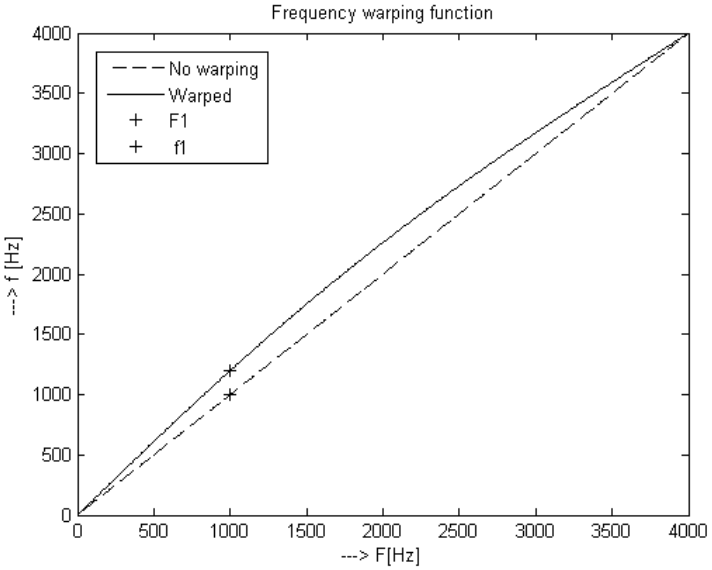


Fig. 5. Frequency warping function with $\alpha = 0.1032$ and $\alpha = 0$

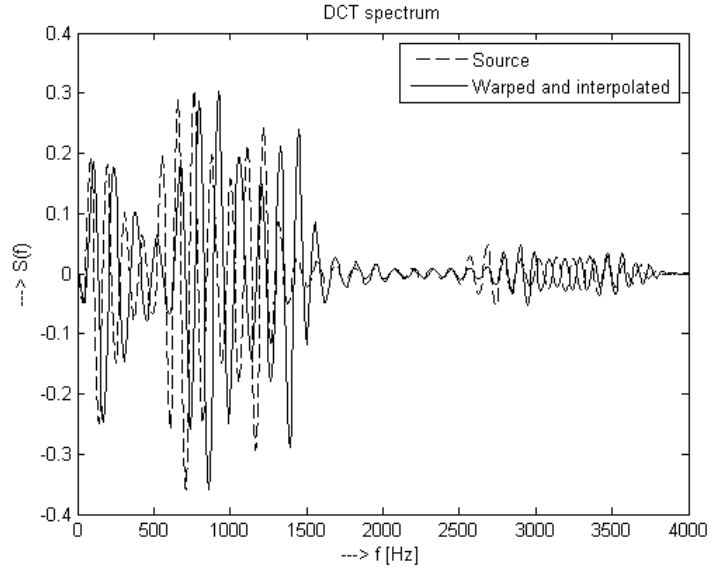


Fig. 6. Original and warped DCT spectra

4 Speech Reconstruction from the DCT Speech Model

The finite impulse response of the converted DCT model is obtained by the inverse DCT (2). The windowed speech signal together with the converted impulse response obtained by the inverse DCT transform for our example are given in Fig. 7.

The converted impulse response is used as the FIR vocal tract impulse response in the speech production model in Fig. 1. It is used pitch synchronously with half frame overlapping (the frame shift is equal to one pitch period). Owing to the fact that the dimension of the DCT $M \geq N$, where N is the frame length, the length of the converted impulse response may be different from N. This can be respected in the pitch synchronous overlap and add construction of the resulting speech. In the case of unvoiced sounds the frame length N is chosen equal to the last voiced frame.

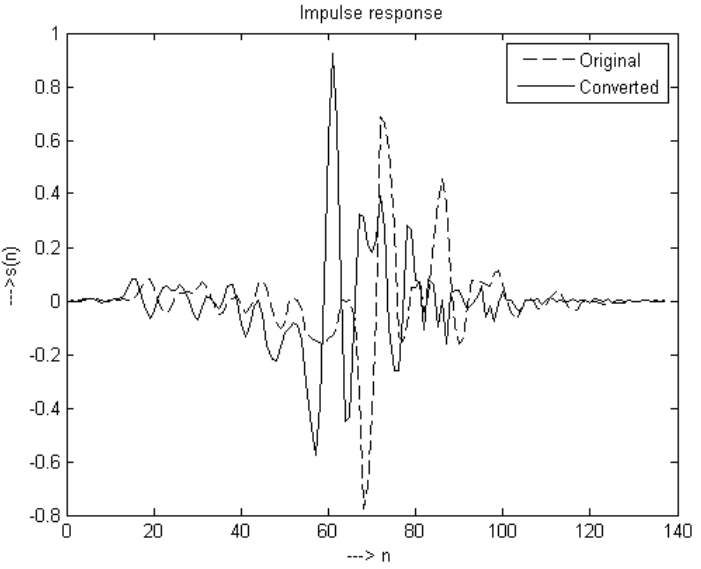


Fig. 7. Windowed speech signal and the converted impulse response

In addition to the speech spectrum warping described in Chapter 3 also the suprasegmental parameters, i.e. the fundamental frequency, the speech rate and the speech intensity of the source speaker can be modified to mimic that of the target speaker. This can be included into the pitch synchronous construction of the resulting converted speech using the procedure proposed in [14]. For the modification of the fundamental frequency and the speech intensity of the source speaker into that of the target speaker a simple mean-variance conversion method is used. For modification of the speech rate we use three speech rate modification factors: for voiced speech, unvoiced speech and for pauses. These factors can be obtained from the analysis of natural speech data.

The whole procedure of DCT voice conversion can be represented by Fig. 8. By comparing this figure with Fig. 2 it can be seen that the quantization and inverse quantization in transform coding are replaced by nonlinear frequency warping in voice conversion. Therefore we called the voice conversion *transform warping*.

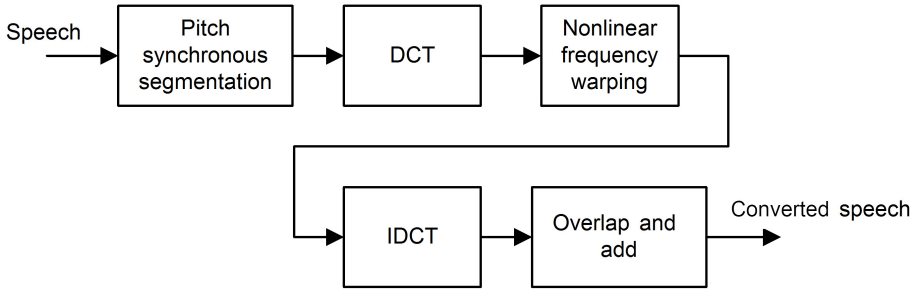


Fig. 8. The flowchart of the DCT voice conversion

5 Conclusion

In this contribution a voice conversion based on frequency warping of the DCT spectrum was presented. The approach is characterized by easy implementation with less computational requirements than FFT-based conversion or using the complex cepstrum implementation. The aim of this method is the generation of new voices for TTS speech synthesis without the generation of new inventories for new speakers.

The proposed voice conversion is based on only one warping function for the whole utterance, which is estimated from the average positions of the formants for the source and target speakers. The quality of the reconstructed speech is thanks to the mixed phase speech modeling very high. The change of the voice identity is reliably reached, but the total similarity of the transformed speech to that of the target speaker is not perfect. The computational requirements are in comparison to the existing voice conversion algorithms very low.

Acknowledgment. This paper has been supported within the framework of COST2102 by the Ministry of Education, Youth and Sports of the Czech Republic, project number OC08010 and by the Grant Agency of the Czech Republic, research project 102/09/0989.

References

1. Moulines, E., Sagisaka, Y.(eds.): Voice Conversion: State of the Art and Perspectives. Special Issue of Speech Communication 16(2) (1995)
2. Kain, A.B.: High Resolution Voice Transformation. PhD Thesis, Oregon Graduate Institute of Science and Technology (2001)

3. Přibilová, A., Přibil, J.: Non-linear Frequency Scale Mapping for Voice Conversion in Text-To-Speech System with Cepstral Description. Speech Communication 48(12), 1691–1703 (2006)
4. Vondra, M.: Voice Transformation in Vocoders and TTS Systems. PhD Dissertation, Brno University of Technology (2005) (in Czech)
5. Nemsak, S.: Pitch Shifting and Voice Transformation Using PSOLA. In: Vich, R. (ed.) Proc. of the 13th Czech-German Workshop on Speech Processing, Prague, September 15-17, pp. 38–41 (2003)
6. Vondra, M., Vích, R.: Speech Identity Conversion. In: Chollet, G., Esposito, A., Faúndez-Zanuy, M., Marinaro, M. (eds.) Nonlinear Speech Modeling. LNCS (LNAI), vol. 3445, pp. 421–426. Springer, Heidelberg (2005)
7. Vondra, M., Vích, R.: Speech Modeling Using the Complex Cepstrum. In: Esposito, A., Esposito, A.M., Martone, R., Müller, V.C., Scarpetta, G. (eds.) COST 2102 Int. Training School 2010. LNCS, vol. 6456, pp. 324–330. Springer, Heidelberg (2011)
8. Vondra, M., Vích, R.: Modification of the Glottal Voice Characteristics Based on Changing the Maximum-Phase Speech Component. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 240–251. Springer, Heidelberg (2011)
9. Vich, R.: Pitch Synchronous Linear Predictive Czech and Slovak Text-to-Speech Synthesis. In: Proc. of the 15th International Congress on Acoustics, ICA 1995, Trondheim, Norway, vol. III, pp. 181–184 (1995)
10. Vich, R.: Cepstral Speech Model, Padé Approximation, Excitation and Gain Matching in Cepstral Speech Synthesis. In: Jan, J. (ed.) BIOSIGNAL 2000 VUTIUM, Brno, pp. 77–82 (2000)
11. Oppenheim, A.V., Schafer, R.W., Buck, J.R.: Discrete-Time Signal Processing. Prentice Hall, New Jersey (1999)
12. Zelinski, R., Noll, P.: Adaptive Coding of Speech Signals. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-25(4), 199–309 (1977)
13. Tribolet, J.M., Crochiere, R.E.: Frequency Domain Coding of Speech. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-27(5), 512–530 (1979)
14. Vondra, M., Vích, R.: Speech Emotion Modification Using a Cepstral Vocoder. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) COST 2102 Int. Training School 2009. LNCS, vol. 5967, pp. 280–285. Springer, Heidelberg (2010)

ATMap: Annotated Tactile Maps for the Visually Impaired

Limin Zeng and Gerhard Weber

Technische Universität Dresden
Institute of Applied Computer Science
D-01062 Dresden, Germany
{limin.zeng,gerhard.weber}@tu-dresden.de

Abstract. For the visually impaired, there are various challenges to understand cognitive spatial maps, specifically for the born blind. Although a few existing audio-haptic maps provide possibilities to access geographic data, most of them are hard to offer convenient services through interactive methods. We developed an interactive tactile map system, called ATMap. It allows users to create and share geographic referenced enhancing annotations on a 2D tactile map, in order to obtain more about relevant places beyond static information in GIS database. 5 blind users have been recruited to evaluate the system in a pilot study.

Keywords: Braille display, tactile map, geographic annotation.

1 Introduction

Maps as one of the most important mobility tools are available on printed paper, various computer-based devices, and other materials. However, for millions of people who are blind or visually impaired it's hard to access the spatial information from maps and relevant applications. Although there are a number of specific map tools [1], most of them fail to satisfy the users' requirements. For example, tactile maps on raised paper only render a small area with reduced information. Single line Braille displays and speech-based screen reader software are unable to represent the spatial relationship of geographic data precisely. A suitable approach to represent the geographic data, and design non-visual interaction for the visually impaired is needed.

In recent years, a couple of tactile map systems have been developed to improve the accessibility of geographic maps, and to support inquiring names of geographic features. Beyond the geo-data from GIS servers, however, the disabled users expect to obtain dynamic and detailed descriptions of the points of interest (POIs) in order to satisfy their demands and to learn about the latest changes or to plan a new route path, which will support decision making.

Aiming at satisfying user requirements and exploring maps easily, in this paper we demonstrate a more interactive tactile map for the visually impaired in a desktop environment. We developed an annotated tactile map system namely ATMap. The ATMap allows users not just to pan, zoom and search on a graphic-enabled Braille display, but also to create and share annotations of points of interest on the maps

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 290–298, 2012.

through a collaborative approach. The results of an evaluation towards familiar and unfamiliar places indicate the blind subjects would access maps and share annotations through ATMap.

2 Background and Related Work

The early printed tactile maps normally produced by thermoform, swell paper or embossed paper are popular for the visually impaired, but they are still unable to satisfy the increasing user requirements, because of rendering less information in a small area. The virtual tactile map was developed to provide more information and flexible methods to explore maps through a computer [3]. However, the virtual tactile maps are hard to help users build up a spatial cognitive map due to lack of explicit representation of maps. For instance, users would hardly understand the precise orientation towards a bus stop or the distance between two POIs on virtual maps. Audio-haptic interaction with force-feedback devices on virtual maps may provide more details [4] but is difficult to use for zooming.

With the purpose of integrating advantages from printed tactile maps and virtual tactile maps, a mixed solution has been proposed to offer much more information by combining audio output and directly touching a real raised map, which is mounted on a touch-sensitive panel ([5] and [6]). Nevertheless, the above methods are also restricted by the limitations of the printed map size and the manual production. In recent years, a couple of graphics-enabled Braille displays [7][8] and touch-screen devices have been designed and enhance access to graphical information. People who are blind or visual impaired not only are enabled to use computers in a GUI window [9], but can also explore digital city maps through searching, panning and zooming by finger-gestures on the surface [2]. This kind of tactile maps completely overcomes the limitation of the size, and provides a variety of interactive approaches to let users read maps more conveniently.

Furthermore, beyond the static geographic data stored in a GIS server, users need more detailed and dynamic updates and annotations from themselves or others. User-generated geographic data benefit the sighted, like OpenStreetMap [10], and would benefit the disabled as well [11] and [12]. Even if people are able to annotate through GPS-enabled mobile devices while walking outdoor [13],[14], for the visually impaired it's hard to learn spatial mental maps through relevant annotations on the virtual maps. However, there is no system allowing them to create and share annotations via a real tactile map, which renders the relationship of geographic features explicitly.

3 An Annotated Tactile Map

The new touchable annotated tactile map system (ATMap) allows the visually impaired to make annotations on geographic features. Geographic annotations on ATMap are text linked to some geographic position and include comments aiming at users who are blind. The tactile map employs a multi-line desktop Braille display (array of 60x120 pins) to render maps via the raised pins against finger-tips. Due to its

touch-sensitive surface, the display enables users to operate maps by finger-gestures, like zooming, panning. In order to interact with maps more conveniently while creating and sharing annotations, several functions are necessary, e.g. map overview, Braille caption, and search nearby POIs by touching.

As illustrated in Figure 1, there are several main modules in the system. The server contains map data and annotation data from users, and the annotation module allows users to write and read geographic referenced annotations. The exploration module responses to users' basic gesture commands (e.g., tapping, panning and zooming), as well as to other functions. The map presentation module will render various map elements via tactile map symbols on the multi-line Braille display. The Braille display is connected to a computer via USB interface, and typically refreshes at every 1 Hz.

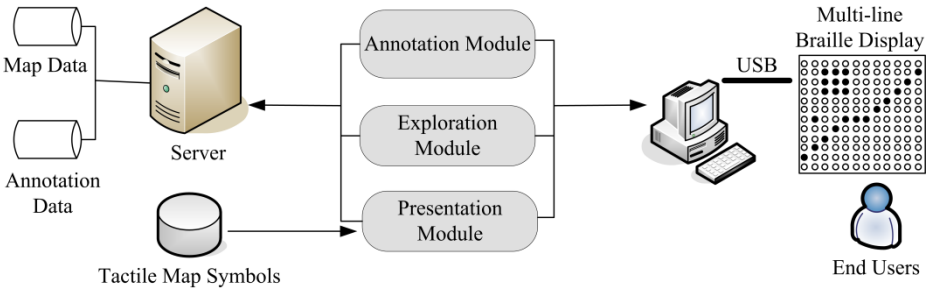


Fig. 1. The system structure of ATMap

3.1 Tactile Maps on Multi-line Braille Display

As one of important components, map data are stored and delivered by a mainstream GIS server. ATMap obtains map elements through standard Geography Markup Language (GML)[1] from the GIS server in particular street names or landmarks. With the help of this approach, no images have to be processed for tactile presentation. Even in a complex area detailed attributes of map elements may be obtained, including geographic regions and their categories.

Refreshable Braille displays only allow representing information by raised or lowered pins, which is different to rendering maps in a visual channel by colors, text and even overlapping layers. In order to represent an accessible tactile map with as many map elements as possible, we design a set of tactile symbols through raised pins (see Figure 2). Users can identify streets, buildings and various POIs.

Due to the touch-sensitive surface, multitouch interaction is available [15] on the device. ATMap recognizes one finger tap over tactile symbols to obtain related attributes, one finger moving for panning and two fingers pinching for zooming operations. Furthermore, the system updates the whole tactile display automatically after panning or zooming.

[1] GML is an XML grammar for rendering geographic features, and is an ISO standard `http://www.opengeospatial.org/standards/gml`

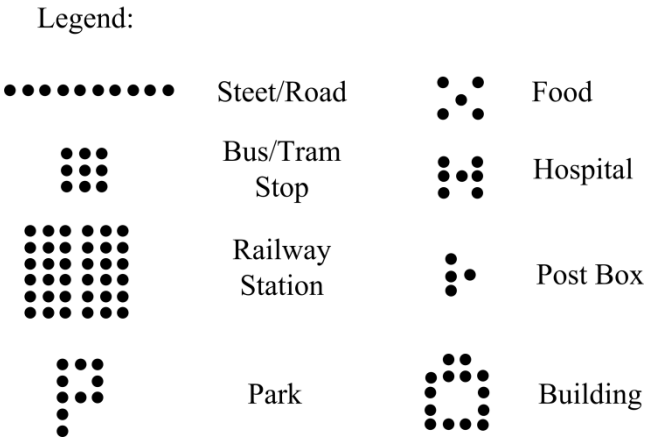


Fig. 2. Examples of tactile map symbols (black points mean raised pins)

3.2 Map Overview and Braille Caption

Since it's a time-consuming task for the visually impaired to explore the tactile display and read geographic features rendered on the map, users may query elements according to specific categories, e.g. streets, public transportation, buildings, or POIs. The area of the map is divided into 6 equal zones (2×3 array, see Figure 3), indicating in which zone the relevant geographic features are. For instance, as shown in Figure 3, when users want to know tram stops on the tactile display, the system informs them about the name and the zone as "Münchner Platz in Zone 4".

A Braille caption window at the bottom of the display renders alternative descriptions in Braille apart from the audio output. Thus, users are able to choose the way to obtain information by audio, in Braille or both. Such Braille captions may even be accessible to deaf-blind users.

3.3 Nearby POI Search

Through GPS-enabled devices the visually impaired find out nearby POIs easily and convenient. However, in the desktop configuration the traditional approach is by searching POIs via keywords. Due to the touch sensitive surface, we developed a novel function to search nearby POIs through tapping the map with one finger. This method only needs to input a selected radius by keyboard input, and is able to find more nearby POIs without knowing their names or categories in advance. Furthermore, users understand the detailed spatial positions of POIs and their spatial relationships on the large tactile map. For instance, Figure 4 illustrates a scenario showing a blind user's search for the vicinity nearby to her/his new house. The users types "500 meters" by a keyboard to seek POIs in 500 meters distance from the origin as specified by click gesture.

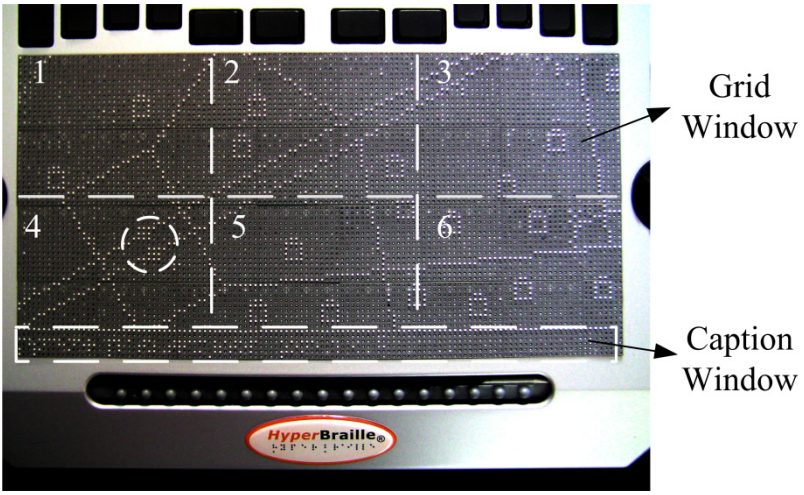


Fig. 3. A grid layout with 6 zones (2×3 array) and a Braille caption window (the circle is the tram stop, namely Münchner Platz)*

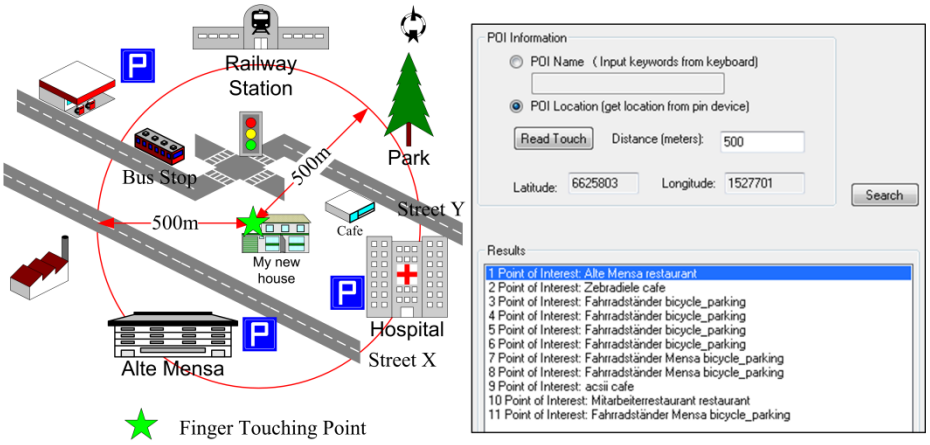


Fig. 4. What's near to my new house in 500 meters (left: the scenario on the map; right: the search result dialog to show nearby POIs)?

3.4 Maps towards Enhancing Annotations

Individuals with visual impairments read geographic data like street names or stations from existing tactile maps which are based on geographic information systems (GIS). Nevertheless, they expect more detailed annotations within those places, as well as

* Note: the real color of its surface is black, and the pins are in white, see `http://www.hyperbraille.com/`, and the dash lines and numbers as markers are drawn on the original image.

relevant suggestions from other users and being able to share their experiences. In contrast to the geographic data, the user-generated annotations are created from a personal perspective and remind their creators in the future. These annotations might be interesting to other users as well. Therefore, all of the users will benefit from annotations before or during their journey.

A collaborative user interface allows users to create and share annotations for POIs on a map. A central server stores and distributes relevant annotations. Figure 5 represents in a scenario, how a new blind resident learns from previous annotations about a tram stop, and becomes careful at crossings due to traffic lights without audio output. The user is not only benefitting from the location-based annotations, but is also encouraged to submit her/his experiences to service others through this location-based service.

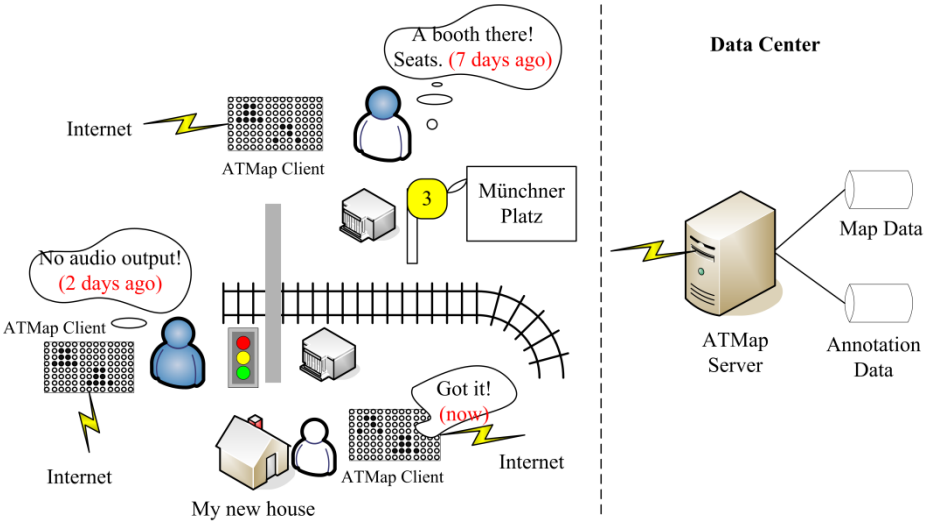


Fig. 5. A collaborative annotation model to share updated information from system server on ATMap clients

4 A Pilot Study

4.1 Participants

We recruited 5 legally blind subjects (4 blind and 1 low vision, aged between 29 and 50, mean age 37) to evaluate the accessibility of creating and reading annotations on the tactile map. One of them is a college student, and others are social workers. The subject with low vision became severely visually impaired about two years ago, and was not familiar with Braille displays yet.

4.2 Procedure

The map data is downloaded from the site of OpenStreetMap[2], and covers 3 city centers with about 18,000 POIs in Germany. Before the test, each subject received training on the handling of the tactile map system, and finished a training task to create and read annotation on a specific POI. When all of the related skills have been understood well, subjects are asked to conduct formal tasks, in which all participants need to read and write annotations on several pre-selected POIs. The pre-selected POIs have been divided into 2 groups through inquiring subjects before the formal tasks, where one group contained POIs known or visited before, and the other group was unknown.

During the formal tasks, the participants had to find out candidate POIs (at least one unknown POI and one known POI) by keyword searching and undertake a nearby search. Each task includes then to create and read annotations. While searching by keywords, the subjects have to use a computer keyboard to input the name of POIs through assistive software (e.g. a screen reader). When using the method of nearby searching the subjects need to decide about a radius and tap on the surface, as shown in Figure 4. Subsequently the subjects select the candidate POIs from the results obtained and shown tactile. Finally, the subjects input text annotations in a popup window through the screen reader. After the evaluation, a post questionnaire is completed to obtain users' feedback and suggestions.

4.3 Results

At the end of the evaluation, we collected 25 annotations in total on the tactile display, of which 5 were from training tasks and remaining ones from formal tasks. The subjects wrote down a text description on their experiences about a bus stop, favored food in a restaurant, or question about an unvisited place. All of the subjects were able to read and create annotations in a short time (1-2 minutes) successfully. Errors happened while searching via keywords, as the subjects had to correctly spell the names of POIs whose spelling was unfamiliar to them, specifically for the unknown POIs. There were also mistakes made when inspecting candidate POIs determined by nearby search. When the value of the radius is too large, some subjects had to spend more time to choose the correct POIs from a long list of results.

The four blind participants stated that they preferred nearby searching over keyword searching. They reported they would obtain the detailed locations of POIs on the map explicitly, which helped them to remember the spatial relationships of POIs. Furthermore, all of the participants were interested in the collaborative approach to share location based enhancing annotation, because they would not only learn updated geographic data, but also would obtain valuable experiences from other users to support them when making decisions. The low vision subject liked to provide keyword input on the screen with the help of magnifier software, but she explained that it was also an impressive interaction on the tactile display.

[2] www.openstreetmap.org

4.4 Discussion

Currently, when searching POI keywords on map applications like Google Map, the visually impaired access text descriptions from the result lists, rather than related locations. The proposed nearby POI search would let them determine the spatial relationship. However, we observed that the size of the specified radius is a significant factor to obtain a meaningful result set. When the size is too large, there are too many items. To refine the results, it might be a better solution to employ the orientation to the contact point as another parameter in nearby POI search.

In addition to various novel interactive human computer interfaces, the emerging social interaction should be one of the new features towards enhancing usability and accessibility of assistive technology for people with special needs. Visually impaired people not only share geographical referenced annotations to others, but also obtain help from someone who is personally unknown. Providing experiences to others may lead to new opportunities for higher mobility of blind people in general.

Furthermore, it's hard to collect annotations only by the disabled users as detailed and as many as possible, thus, a group of volunteers should be encouraged to share their experiences with the disabled. A possible method is to connect the system to the mainstream social networks, like Facebook and Twitter. However, the volunteers may do not know the different concerns for the disabled with different special needs. For example, the wheelchair users need to know which route has no stairs, and the blind users expect a traffic light with audio feedback. It's necessary to define the meta-data of annotations for different user groups by an expert, who understands user requirements. As the amount of annotations sharply increase, another issue is how to manage the temporal properties of annotations. At the same time, the issue of privacy requires more work to integrate some method of building reputation when relying on social interaction.

5 Conclusion and Future Work

The development of touchable screen devices and emerging multi-line Braille displays, offer various novel user interfaces and more interactive experiences to offer new services for the visually impaired, specifically in accessing graphics-based information. In this article, the authors developed several new approaches to improve accessibility to tactile maps on a touchable multi-line Braille display, in which the function of nearby POI searching via a finger helps users to find out surroundings in an unfamiliar environment easily, and may reduce cognitive load.

More importantly, the traditional geo-data in GIS are not enough dynamic and personalized in today's digital world, not only for the sighted people but also for the visually impaired people. Annotated tactile maps are an initial step to collect and share dynamic geo-referenced annotations. 5 visually impaired participants joined the evaluation, and all of them access map data with the new user interfaces and, furthermore, they share and read geo-referencing annotations.

For the next steps, due to limited the number of participants in the current pilot study, the authors plan to move further into the evaluation, in order to investigate

users' cognitive maps when interacting through panning and zooming on the multi-line Braille displays. Because the visually impaired currently use ATMap for their pre-journey in a desktop environment, we are interested in implementing a mobile version in the future, to access interactive tactile maps on the move.

Acknowledgements. We thank our participants for the comments on ATMap system, and appreciate supports from China Scholarship Council (CSC).

References

1. Zeng, L., Weber, G.: Accessible Maps for the Visually Impaired. In: Proceedings of IFIP INTERACT 2011 Workshop on ADDW, CEUR, vol. 792, pp. 54–60 (2011)
2. Zeng, L., Weber, G.: Audio-Haptic Browser for a Geographical Information System. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010, Part II. LNCS, vol. 6180, pp. 466–473. Springer, Heidelberg (2010)
3. Schneider, J., Strothotte, T.: Constructive Exploration of Spatial Information by Blind Users. In: Proceedings of Assets 2000, pp. 188–192 (2000)
4. Springsguth, C., Weber, G.: Design Issues of Relief Maps for Haptic Displays. In: Proceedings of HCI International 2003, vol. 4, pp. 1477–1481 (2003)
5. Miele, J.: Talking TMAP: Automated Generation of Audio-tactile Maps Using Smith-Kettlewell's TMAP Software. British Journal of Visual Impairment 24(2), 93–100 (2006)
6. Wang, Z., Li, B., Hedgpeth, T., Haven, T.: Instant Tactile-audio Map: Enabling Access to Digital Maps for People With Visual Impairment. In: Proceedings of Assets 2009, pp. 43–50 (2009)
7. Kraus, M., Völkel, T., Weber, G.: An Off-Screen Model for Tactile Graphical User Interfaces. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 865–872. Springer, Heidelberg (2008)
8. KGS DotView tactile graphic display, http://www.kgs-jpn.co.jp/b_dv2.htmlHH (last accessed in April 2012)
9. Prescher, D., Weber, G., Spindler, M.: A Tactile Windowing System for Blind Users. In: Proceedings of Assets 2010, pp. 91–98 (2010)
10. Haklay, M., Weber, P.: OpenStreetMap: User-generated Street Maps. IEEE Pervas. Comput. 7(4), 12–18 (2008)
11. Rashid, O., Dunbar, A., Fisher, A., Rutherford, J.: Users Helping Users: User Generated Content to Assist Wheelchair Users in an Urban Environment. In: Proceedings of 2010 9th Inter. Conf. on Mobile Business/2010 9th Global Mobility Roundtable, pp. 213–219 (2010)
12. Völkel, T., Weber, G.: RouteCheckr: Personalized Multicriteria Routing for Mobility Impaired Pedestrians. In: Proceedings of Assets 2008, pp. 185–192 (2008)
13. Espinoza, F., Person, P., Sandin, A., Nyström, H., Cacciatore, E., Bylund, M.: GeoNotes: Social and Navigational Aspects of Location-Based Information Systems. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) UbiComp 2001. LNCS, vol. 2201, pp. 2–17. Springer, Heidelberg (2001)
14. Burrell, J., Gay, G.: E-graffiti: Evaluating Real-world Use of a Context-aware System. Interacting with Computers 14, 301–312 (2002)
15. Schmidt, M., Weber, G.: Multitouch Haptic Interaction. In: Stephanidis, C. (ed.) UAHCI 2009. LNCS, vol. 5615, pp. 574–582. Springer, Heidelberg (2009)

From Embodied and Extended Mind to No Mind

Vincent C. Müller

Anatolia College/ACT, Pylaia, Greece &
Programme on the Impacts of Future Technology, University of Oxford, UK
vmueller@act.edu
www.sophia.de

Abstract. The paper will discuss the extended mind thesis with a view to the notions of "agent" and of "mind", while helping to clarify the relation between "embodiment" and the "extended mind". I will suggest that the extended mind thesis constitutes a *reductio ad absurdum* of the notion of 'mind'; the consequence of the extended mind debate should be to drop the notion of the mind altogether – rather than entering the discussion how extended it is.

1 The Standard View

The standard view is that self-contained agents pursue their own goals, sometime in cooperation with other agents, and sometimes using external tools. This typically, but not necessarily, goes together with a view of these agents as rational agents that perceive, then plan and finally act; and the view that robots should be built that way: with sensors, processor and effectors (this view of agents through internal cognitive states, rather than behavioral dispositions, is what I would call 'cognitivism'). This view, in turn goes together with the view that humans and other natural cognitive agents are computational information processors made up of several modules that take in symbolic representations of the world, process these according to specified rules and then produce a symbolic output (this view I call 'computationalism'). In philosophy, cognitivism and computationalism are often take to be scientific explanations of the traditional view that humans have a mind and mental states, and that these states partially explain human behavior. Even those that reject either one or both of these explanations tend to maintain that the traditional view of the mind is largely correct (the Churchlands are a notable exception).

I will suggest that all of the traditional view above is false, but I will argue only for the falsity of the first statement, about self-contained agents and the last, about humans having minds.

2 Embodiment

The rejections of some of the traditional theses take various forms and there is a rather confusing landscape of options. However, one point of criticism is that the traditional view – in distancing itself from its original opponent, behaviorism – puts undue

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 299–303, 2012.

emphasis on a central processing notion of cognition, it talks rather as if cognition was something that I, the agent, do from within my body, taking the information from my sensory system as input, processing this and producing output in the form of actions (typically movements). This image, which has first been properly developed by René Descartes talks as though there were a little man, a homunculus, inside me watching a theatre play – what Dennett aptly called the 'Cartesian theatre' [1]. "… it is a mistake to believe that the brain has any deeper headquarters, any inner sanctum arrival at which is the necessary or sufficient condition for conscious experience." [2]. But not only is there no little man inside me and cognition cannot fruitfully be explained by this model, but the model seems inconsistent, even: The little man would seem to need yet another little man to watch what *he* is doing, etc. – or, in a different terminology, the information would have to be 'encoded' in some way, which results in a need for further decoding of the decoding [3].

Instead of this image, we need to take in many of the cognitive features of the agent that only come into existence due to the interaction with the environment. Also, it seems that any symbols in the cognitive systems need to have "grounding" [4] in physical interaction with the world, in order to be meaningful for the system. The cognitive system is thus embodied in the sense of a "dynamical interaction (coupling) of an embodied system that is embedded in the surrounding environment", "it never goes completely formal" [5]

The rejections of cognitivism and the rejection of computationalism are often lumped into one, presumably because a rejection of cognitivism is thought to imply a rejection of computationalism – but this might not be true (certainly not for pancomputationalists) and the inverse is clearly not true. Descartes was not a computationalist, but he surely was a cognitivist.

In my view, the thesis that "cognition is embodied" takes three main forms, which in turn have their variations:

1. Embodiment as an *empirical thesis*. For example:
   - Sensation and experience require movement (e.g. of eyes or percept), so perception is a kind of action [6][7] and we should really speak of a "sensorimotor system" rater than "sensory system"
   - Conscious experience is action experience [8]
2. Embodiment as an *engineering thesis*. For example:
   - Many tasks can be achieved by active control or by body morphology (e.g. running) [9]
   - Body involvement is a design choice (e.g. active sensing) … [10]
3. Embodiment as a *metaphysical thesis*. For example
   - There can be no disembodied homunculus inside watching a 'Cartesian theatre'
   - There can be no meaningful symbols in a cognitive system without embodiment and embeddedness

I think it will become clear shortly that the extended mind thesis is first and foremost a metaphysical thesis, which then has an empirical consequence (the human mind is often extended) and an engineering consequence (it does not matter where you locate the resources for a cognitive function).

3 Extended Mind

Andy Clark and David Chalmers [11] have proposed the much-discussed thesis that cognitive processes of humans can and do take place outside the head; in particular that artifacts we use, like notebooks or electronic devices are part of our cognitive apparatus. We are thus, in Clark's words "Natural Born Cyborgs" [12] with "Supersized Minds" [13]. I will introduce the 'extended mind' thesis and try to find out which consequences we should draw from the discussion – in particular for the notion of the cognitive 'agent' and for the 'embodiment' of agents.

Clark and Chalmers show a number of examples where it does not seem to matter whether the human cognitive activity takes places 'in the head' or outside: rotating blocks mentally or physically (to see whether they would fit a gap in the computer game 'Tetris'), touching something with hands or a stick, counting in the head vs. with fingers and, finally, Inga and Otto who have the belief that "The Museum of Modern Art is on 53rd Street". Since beliefs are still the staple 'mental state' for most philosophers, this example in [11] has produced the most debate.

Inga knows there the museum is because she remembers it, quite normally. Otto also knows where the museum is, but he has Alzheimer's Disease and thus keeps such information in a notebook that he can consult. If you find the idea of the notebook to 'external', imagine that Otto has a brain implant that functions as his notebook. Also note, that we quite naturally say such things as "I know what time it is" (because I have a watch) or "I believe I have an appointment" (my computer says so).

So, what we have is extended mental processes (like mental rotation), extended perception and extended belief – in short, the extended mind.

The main line of the extended mind thesis is often summarized in what Clark calls a *Parity Principle:*

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head,* we would have no hesitation in recognizing as part of a cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. [11][cf. 13]

This principle is meant to overcome the traditional image which is "… in the grip of a simple prejudice: the prejudice that whatever matters about mind must depend solely on what goes on inside the biological skin-bag, inside the ancient fortress of skin and skull." [14]. Instead, we should accept that "non-biological resources, if hooked appropriately into processes running in the human brain, can form parts of larger circuits that count as genuinely cognitive in their own right" [15] It does not matter

that these processes are not biological and it does not matter *how* they are hooked into the processes – using perceptual apparatus (as in the notebook) is just as acceptable as a more direct brain interface. What matters is that they are intuitively mental, in particular they function as such. So, the extended mind thesis is that mental processes do not only take place inside the skull or skin.

4 Conclusion

If we consider the full picture of 'cognition', we can not restrict ourselves to what is 'inside the skin', we must allow for cooperation, even intelligence of complex wholes (like 'swarms'), for cognitive offloading onto the environment and culture [16], for construction of our own cognitive niche [17][13] and we must remember that much of the abilities of agents are due to the morphology of their bodies [9]. This does not mean, however, that we must conclude that 'the mind is extended' – because that becomes absurd – but that we must forget about describing the abilities of such agents and systems in terms of 'minds' and their location.

Instead, we must admit that out perspectives and explanatory purposes determine where we want to make the 'cut' of what counts as 'one agent' – and then the best explanation wins, whether it involves only systems inside a skin or not. The notion that is left is the "person" – which we need for ethics (but it has no sharp boundaries and is dependent on purposes). What we must do is to forget about 'the mind' – instead, ask 'how does this work? In other words: The proof is in the pudding. It is time to change perspective: The mind is dead.

References

1. Dennett, D.C.: Consciousness Explained. Little, Brown & Co., New York (1991)
2. Dennett, D.C., Kinsbourne, M.: Time and the Observer: The Where and When of Consciousness in the Brain. Behavioral and Brain Sciences 15, 183–247 (1992)
3. Bickhard, M.H.: Representational Content in Humans and Machines. Journal of Experimental and Theoretical Artificial Intelligence 5, 285–333 (1993)
4. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
5. Calvo, P., Gomila, T. (eds.): Handbook of Cognitive Science: An Embodied Approach. Elsevier, München (2008)
6. Noë, A.: Action in Perception. MIT Press, Cambridge (2005)
7. Myin, E., O'Regan, K.J.: Studied Perception and Sensation in Vision and Other Modalities: A Sensormotor Approach (2006)
8. O'Regan, K.J.: Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness. Oxford University Press, New York (2011)
9. Pfeifer, R., Bongard, J.: How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press, Cambridge (2007)
10. Cangelosi, A., Riga, T.: An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments with Epigenetic Robots. Cognitive Science 30(4), 673–689 (2006)
11. Clark, A., Chalmers, D.J.: The Extended Mind. Analysis 58(1), 7–19 (1998)

12. Clark, A.: Natural Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. Oxford University Press, Oxford (2003)
13. Clark, A.: Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford University Press, New York (2008)
14. Clark, A.: Natural Born Cyborgs? Edge (December 28, 2000), http://www.edge.org
15. Clark, A.: Letter on Fodor on 'Where Is My Mind? London Review of Books 31(6) (March 26, 2009)
16. Hutchins, E.: Enculturating the Supersized Mind. Philosophical Studies 152(3), 437–446 (2011)
17. Clark, A.: Language, Embodiment, and the Cognitive Niche. Trends in Cognitive Sciences 10(8), 370–374 (2006)

Effects of Experience, Training and Expertise on Multisensory Perception: Investigating the Link between Brain and Behavior

Scott A. Love[1], Frank E. Pollick[2], and Karin Petrini[3]

[1] Indiana University, Department of Psychological and Brain Sciences, Bloomington IN 47405 USA
[2] University of Glasgow, School of Psychology, 58 Hillhead Street, Glasgow G12 8QB UK
[3] University College London, Institute of Ophthalmology, UK
sclove@indiana.edu, frank.pollick@glasgow.ac.uk, k.petrini@ucl.ac.uk

Abstract. The ability to successfully integrate information from different senses is of paramount importance for perceiving the world and has been shown to change with experience. We first review how experience, in particular musical experience, brings about changes in our ability to fuse together sensory information about the world. We next discuss evidence from drumming studies that demonstrate how the perception of audiovisual synchrony depends on experience. These studies show that drummers are more robust than novices to perturbations of the audiovisual signals and appear to use different neural mechanisms in fusing sight and sound. Finally, we examine how experience influences audiovisual speech perception. We present an experiment investigating how perceiving an unfamiliar language influences judgments of temporal synchrony of the audiovisual speech signal. These results highlight the influence of both the listener's experience with hearing an unfamiliar language as well as the speaker's experience with producing non-native words.

Keywords: multisensory, audiovisual, perception, expertise, drumming.

1 Introduction

Everyday we receive a large amount of sensory information, the majority of which is redundant. Some of this information comes from the same source and needs to be combined, whilst other information needs to be kept separate because it arises from a different source. The ability of our brain to process multisensory information and make sense of it is essential for our wellbeing and for conducting everyday tasks. Many internal and external factors can dictate whether two sensory signals will be integrated and how. For example, in a situation in which sound localization is limited (e.g. when walking in a very noisy street) combining sound and sight can help us make better decisions and keep us from harm (e.g. when the pedestrian traffic light starts beeping and concomitantly turns green). Thus combining multiple cues can reduce uncertainty and enhance our ability to make better estimates of the situation

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 304–320, 2012.

[1, 2]. Because of this realization, in cognitive neuroscience we have seen somewhat of a paradigm shift away from trying to explain human perception by individuating and understanding each of our senses separately. Indeed the field has moved towards a more holistic approach that considers the interaction between the senses to be, at least, as important as unimodal perception. Shadowing this shift, we begin this article by briefly describing behavioral, functional and neuroanatomical evidence of experience-dependent plasticity of unimodal and multimodal processing. Subsequently, we outline a relatively fresh research strand aiming to specifically understand how expertise can enhance, or fine tune, and alter *multisensory processing*; in particular, how musical training or experience with a particular language can influence audiovisual synchrony perception.

2 Effects of Experience on Unimodal and Multimodal Processing

Several behavioral studies have now demonstrated that experience can modify sensitivity to unimodal sensory information such as vision, touch and sound [3-5], as well as the way we integrate this information [6, 7]. For example, Green and Bavelier [3] reported that playing action video games enhances the spatial resolution of visual processing, and Atkins, Fiser and Jacobs [6] that experiencing haptic information can modify observers' reliability estimates of visual cues during three-dimensional visual perception. This multisensory malleability is not confined to visual-haptic interaction, but extends to audiovisual temporal processesing. Powers, Hillock [8], for example, found that multisensory perceptual training can decrease our tolerance towards audiovisual asynchrony, by reducing the size of the temporal integration window (TIW).

When we integrate information from our various senses, experience matters not only because it can affect the way we estimate cue reliability and subsequently combine the cues, but also because it can affect the role of prior knowledge in cue combination. Prior knowledge stems from previous experience of the world and in certain instances it may be innate [9-13]. A striking example of prior knowledge that is often reported is an illusion called the 'hollow-face' [14] in which a concave mask elicits the percept of a convex face. This happens because the prior belief that faces are convex overrides the sensory information of concavity. In an elegant study Adams, Graf [15] also showed that the prior assumption of light-from-above can be changed by repeated haptic feedback and that this adaptative mechanism extends to different situations and tasks. These findings clearly suggest that our behaviors and assumptions are not only dependent on sensory information but are constantly shaped by prior experience.

Along with human behavior, brain structure and function are remarkably plastic; moreover, recent reviews have outlined the general nature of the experience-dependent organization of both cortical and sub-cortical brain regions [16-18]. These reviews highlight that neuroanatomical organization and behavior can be modified by many different types of expertise as well as by learning over both long and short

periods of time. For example, learning the identity of unfamiliar voices over a short period, i.e., around 6 twenty minute learning sessions, produces significantly improved vocal identity discrimination performance and alters how voices are processed in the inferior frontal cortex [19]. The N170 electrophysiological component, most often referenced in regards to face processing [20], is also modulated by expertise with objects other than faces such as 'greebles' [21, 22], dogs and birds [23], and fingerprints [24, 25]. Effects of expertise on brain activity have also been reported when using functional magnetic resonance imagining (fMRI) to compare expert dancers to non-experts [26-28]. For example, Calvo-Merino and collaborators found that expert dancers and non-experts differed in the level of activation in fronto-temporal-parietal cortex when viewing dancing actions [26]. That is, dancers had greater activation for movements that they had been trained to perform than for those that they had not, while non-experts did not display any difference, demonstrating that the specific motor expertise was key to explaining changes in brain responses.

In regards to brain structure, the length of time London taxi drivers have been in their job correlates with how much larger their posterior hippocampi are than those of controls [29, 30]. Similarly, the posterior hippocampi of dancers and slackliners are larger than those of controls [31]. Moreover, the gray matter volume of several cortical areas are known to increase after just 40 hours of golf training [32]. These are just a few examples of how learning and expertise in various areas can produce skilled behavior that can be associated with plastic reorganization of brain structure and function. Several studies have now shown that inter-individual variation in white matter reflects behavioral variation [33, 34]; however, they were not able to determine a causal role of experience or training on white matter structures. However, the existence of such a causal role on grey matter changes was demonstrated in 2004 by Draganski and collaborators [35]. In a longitudinal MRI study, the authors compared changes in grey matter structures in a group of adults that were trained to juggle (a motor skill that requires accurate bimanual movements) to a group of non-jugglers. Jugglers and non-jugglers were scanned before and after training and using voxel-based morphometry (VBM) Draganski and collaborators showed an enhancement in grey matter in a mid-temporal area (hMT/V5) and in the left posterior intraparietal sulcus of the jugglers after the training. In a more recent longitudinal study this evidence of a causal relationship between training and changes in brain structures was extended to white matter [36], by using a similar task and diffusion tensor imaging (DTI). Although it cannot be excluded that in some instances experts naturally have larger brain structures, the aforementioned findings provide strong evidence for the direction of causation between changes in behavior and changes in brain structure and function. Nevertheless the real extent and causal relationship of training-induced brain changes can only be completely understood by studying skilled performers after long-term training. This is one reason why musicians are a very useful group to study brain plasticity and the neural correlates of skilled performers.

Differences between musicians and non-musicians in both gray and white matter have also been consistently outlined [37-45]. For example, the cerebellum of male professional keyboard players is larger than that of controls [40] and both singers and instrumentalists

have larger white-matter tract volume and fractional anisotropy in bilateral arcuate fasciculi [41]. Interestingly, a review of this literature pointed out that the specific neuroanatomical changes that occur are dependent on the particular domain of musical expertise [46]. Further to these structural differences, [47], used proton MR spectroscopy to highlight differences between musicians and non-musicians in their concentration levels of the N-acetylaspartate metabolite within the planum temporale.

3 Effects of Musical Experience on Unimodal and Multimodal Processing

Over the last two decades musical expertise has been extensively used as a model to investigate brain plasticity [48-52]. Musical expertise is achieved over many years of extensive training, which fine-tunes and enhances perceptual, cognitive and motor abilities. Crucially, musical expertise is also a specialization that not every individual undertakes; this enables researchers to study plasticity by comparing two different cohorts, i.e., musicians and non-musicians, or by longitudinally observing musical novices as they become musical experts [53]. Furthermore, musical training does not only improve musical ability it can also enhance behavioral performance on a variety of other cognitive abilities: speech perception and linguistic ability [54], second language linguistic skills [55], verbal working memory [56], musical and non-musical auditory imagery [57], visuospatial perception and imagery [58] and even mathematical ability [59]. Hence, the musician's brain is not only an ideal model to explore experience-dependent plasticity in regards to musical experience but also on how experience in one domain transfers to others. How musical expertise transfers to speech perception and linguistic ability is arguably one of the most extensively studied areas of experience-dependent plasticity [55, 60, 61].

The extensive work cited above focuses mainly on how musical experience can shape unimodal processing; however, musical experience, in particular, is inherently multisensory and it would therefore be prudent to use this type of training to explore the experience-dependent nature of multisensory processing. Interestingly, the importance of experience for the development of multisensory integration has been observed at the single-neuron level in non-human animal research [62, 63]. For example, when neurons in the superior colliculus of the cat are deprived of input from the cortex these neurons fail to develop the ability to integrate multisensory information [64, 65]. Over the past five years we have been building upon the small amount of research that has used musical experience to explore how multisensory processing can be altered after many years of musical training [66-73].

One of the first studies investigating the effects of musical training on multimodal information processing and cortical plasticity was conducted by Schulz and collaborators [73]. They used magnetoencephalography (MEG) to compare a group of professional trumpet players with a group of non-musicians and showed that the musicians processed multimodal information (haptic/auditory) differently from non-musicians: when the lower lip was stimulated simultaneously with a tone a different response was elicited compared to when either the lip or the tone were stimulated separately. A subsequent study [74] has since extended these findings by showing that even short periods (2 weeks) of multimodal musical training (i.e., playing musical

sequences on piano) can induce cortical brain plasticity and elicit differential responses compared to unimodal training (listening and make judgments about the musical sequences executed by the multimodally trained group). Specifically, the multimodally trained group showed an enlargement of mismatch negativity (MMNm) from magnetoencephalographic measurements after training when compared to the unimodally trained group. The extent of musical training on multimodal processing leading to brain plasticity and reorganization is not limited to the cortical sensory structures and to a specific age. Indeed, modifications elicited by musical training extend to subcortical sensory structures responding to auditory and audiovisual information [72], and can be detected from early childhood [37, 49, 75]

3.1 Effects of Musical Experience on Perception of Audiovisual Synchrony and Congruency

One way that expertise in multisensory processing of music has been studied is to examine differences between novices and musicians [66-70] in their sensitivity, to audiovisual asynchrony [76-82], to audiovisual congruency [83-86], and to the interaction between these two processes [68, 78, 87, 88]. Since audio and visual channels have different processing latencies due to dissimilarity in physical and neural transmission [76, 89-91] the problem of how they are combined to obtain a unitary percept is not trivial. The amount of asynchrony that can be tolerated while still perceiving the audio and visual streams as unitary is known as the "Temporal Integration Window" (TIW). This window gives a good behavioral measure of training-induced changes in the way musicians process multisensory information, and can be usefully linked to changes in brain activation. In a very recent study Lee and Noppeney [82] examined the TIW of 18 pianists and 19 non-musicians when viewing audiovisual videos of either speech or piano actions. They showed, in agreement with previous results [68-70], that musicians were less tolerant to audiovisual asynchrony (had a smaller TIW) than non-musicians, and also that this higher sensitivity was specific to the piano displays. Having ascertained that, Lee and Noppeley then used fMRI to examine the neural correlates of this behavioral difference between pianists and non-musicians and reported enhanced asynchrony effects for musicians in left superior precentral sulcus, right posterior superior temporal sulcus/middle temporal gyrus, and left cerebellum and effective connectivity for music in an STS-premotor-cerebellar circuitry. Based on their findings the authors thus conclude that piano practicing affords an internal forward model that enables more precise predictions of the relative timings of the audiovisual signals. This idea builds upon initial observations of musical conductors having more finely tuned auditory and temporal processes than non-conductors as assessed by using behavioral and fMRI methods [67], and in the next section we will further explore this issue by reviewing a series of studies comparing drummers and non-musicians sensitivity to asynchrony.

3.2 Effects of Drumming Experience on Perception of Audiovisual Synchrony and Congruency

For their studies, Petrini and colleagues chose drumming. These movements were chosen since drumming movements are very visually salient, in contrast to some other

musical instruments, where asynchrony could be much harder to detect. Motion capture data of drummers playing a swing groove [92] were shown as point light displays [93] in combination with a sound synthesized from an impact model using input from the motion data (see description of the display in Figure 1). Point light displays (PLDs) allow one to isolate the effects of perceiving biological motion from contextual factors, and the specific rhythmic pattern of the swing groove provided a perfect simple stimulus to differentiate between novices and experts.

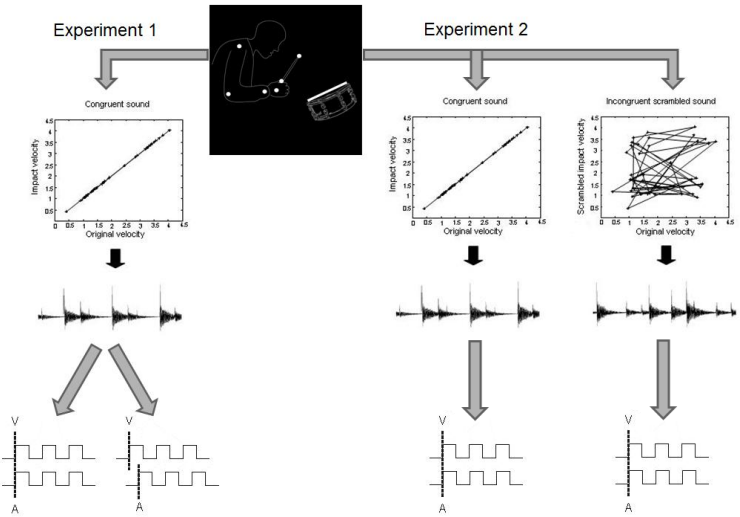


Fig. 1. Schematic of stimulus conditions used in the fMRI study [71]. In the top center of the figure a single frame from the point-light display is presented. The point-light dots represent the drummer's arm beginning at the shoulder joint. Note that the white line outlining the drummer is presented here for clarity only and did not appear in the presented stimulus. For both experiments in left and right columns, the attributes of the visual motion, in terms of the relationship of the original motion velocity relative to implied velocity, appear in the top plots, and the produced sound waveforms appear directly under that. The lowermost panels depict the relationship of the timing of auditory (A) and visual (V) stimuli relative to one another. In Experiment 1 (left column), the displays had an audio that maintained the natural covariation with the visual signal but was presented either in synchrony (left plot on the bottom) or asynchrony (right plot on the bottom). In Experiment 2, the displays had an audio that was always in synchrony with the visual signals, although in one case it covaried with it (left plot on the top) and in the other case it did not (right plot on the top). With kind permission from Elsevier, Petrini et al. (2011) Action expertise reduces brain activity for audiovisual matching actions: An fMRI study with expert drummers. NeuroImage 56, 1480-1492, Fig. 1, Copyright (2011) Elsevier Limited The Boulevard, Langford Lane Kidlington, Oxford, OX5 1GB,UK).

Petrini, Dahl et al. [68] showed that not only are drummers more sensitive to asynchrony (i.e., less tolerant of audiovisual asynchrony), but also that, unlike novices, their sensitivity depends less on the manipulation of other physical

characteristics, such as drumming tempo [68, 77] or audiovisual incongruency [68, 78, 94, 95]. Indeed, while novices are facilitated in detecting asynchrony for drumming displays with faster tempos [68, 69, 77], and also for drumming displays where the covariation between the sound and the drummer's movement has been eliminated [68], drummers are not. The evidence that musicians can tap at slower tempos than non-musicians [96] may explain why drummers are not affected by changes in drumming tempo when judging audiovisual simultaneity. Through practice, drummers acquire the ability to perform drumming actions at a wide range of tempos, which could be why changes in tempo do not affect the way drummers bind the familiar biological motion and its sound. These findings seem to indicate that, after a long period of musical practice, the binding of biological motion and its sound changes in such a way that additional factors are no longer used by our neural system to integrate the multisensory information. This is probably because the system reaches a very high and unbiased level of precision itself, and recent findings seem to further corroborate this conclusion. Petrini, Holt & Pollick [70], for instance, found that only novices' simultaneity judgments were affected by the rotation of a drumming display (point light display in Figure 1 rotated at 90, 180 and 270 degree), while drummers' were not. That is, the tolerance to asynchrony of novices increased when viewing rotated audiovisual drumming displays, while that of the drummers remained relatively unchanged. This extends the findings of Saygin et al. [97], to another kind of audiovisual biological motion event, and indicates that the gestalt of upright point light drumming enhances the detection of audiovisual asynchrony for musical novices but not for expert drummers. Hence, the nature of the visual stimulation can affect the perceived synchrony between the two sensory signals, but the extent of this effect is constrained by the level of experience with a particular multisensory event.

If drummers are better able to detect asynchrony because of their experience and familiarity with that particular biological motion and its resulting sound, then they should still be better than novices when only a part of the body information is presented in the drumming displays. In other words, while the drummers could have acquired, through practice, internal models specific to drumming biological motion that they can use to predict the sound occurrence when no impact point is presented, this should not be the case for the novices. In a further study [69] we addressed this possibility and demonstrated that this is exactly what happens. Not only were drummers found to be better than novices at detecting asynchrony between the drummer's biological motion and the sound, but they were also the only group that could still bind the information from both sensory domains. Indeed it was found that novices were completely unable to discriminate between synchronous and asynchronous drumming displays when the impact point was eliminated. However, when presented with either the intact drumming information (Figure 1) or only the impact point, drummers demonstrated a lack of difference in sensitivity to asynchrony, indicating that as long as the impact point is there they will use it as much as the novices, although maintaining a narrower audiovisual temporal integration window. Thus, while drummers can use both kinds of information, novices can only refer to the impact point when deciding whether or not the sound and the drummer's movement are part of the same action. These findings suggest that

expertise with a certain action enhances the ability to maintain a coherent representation of the multisensory aspects of biological motion. This assumption is strengthened by the finding that when drummers judged the simultaneity between the drumming biological motion and the sound of the aforementioned display from which the impact point information was eliminated, their results were reminiscent of tapping tasks [98, 99], indicating that the acquired information for that specific action was used. In other words, when presented with only the point light arm information of the drumming display, drummers' points of subjective simultaneity occurred in some instances when the sound was leading the sight, showing the same anticipatory effect as that found in tapping tasks [98, 99]. This interpretation suggests that drummers do not possess only a general enhanced ability to determine the co-occurrence of the auditory and visual information for any kind of multisensory event, but they also possess a more specific ability to use the representation of that action to bind sight and sound.

These examinations of the temporal integration window of drummers appear to show that the narrow tuning for audiovisual asynchrony exhibited by the drummers [68-70] and potentially also the ability to fuse sight and sound from incomplete visual displays [69] results from both involvement of higher order (cognitive) processes for the novices in fusing together the audio and visual tracks as well as enhanced perceptual and simulation processes of the drummers in detecting asynchronous events. The neural basis of these differences was studied using brain imaging techniques [71]. Specifically, we used functional Magnetic Resonance Imaging (fMRI) to measure the brain activity of a group of drummers and novices while they watched synchronized and asynchronized PLD drumming displays. Their task was to determine whether the biological motion of the drummer and the sound matched or not (see Figure 1 for a detailed description of stimulus and design used in the fMRI experiments). The timings for the synchronous and asynchronous displays were determined separately for each participant in a behavioral experiment immediately prior to entering the MRI scanner. This predetermination of the optimal timings was necessary to exclude any difference in brain processes between drummers and novices that could be due to differences in task difficulty, rather than in the multisensory processing. Behavioral results from subjects in the scanner indicated that both groups were almost perfect in detecting when the drummer's movement and corresponding sound mismatched in Experiment 1; despite this, the patterns of activation in the detected brain regions were obviously different between the groups [71]. For example, during the task there was a reduced overall activation in bilateral middle frontal gyrus (MFG) for experts compared to novices. Moreover, there was an interaction effect in both the cerebellum and the parahippocampal gyrus: drummers activated these regions less than novices but only to the synchronous stimulus displays. In line with these findings, drummers were found to have reduced activation in fronto-temporal-parietal regions in Experiment 2 were the congruency between the drummer's movements and resulting sound was manipulated (in Experiment 2 sound was always synchronized with the drummer's movements, but the natural covariation between sound intensity and velocity of the drumming strike was manipulated). Our results are complementary to those of Lee and Noppeney [82] in showing finer tuning in musicians than non-musicians when processing audiovisual synchrony information.

Indeed, not only did we find differences in musicians and non-musicians in similar areas (e.g., cerebellum and precentral gyrus), despite examining two different types of musical expertise (piano and drum players), but also whereas Lee and Noppeney [82] show enhanced asynchrony effects for musicians, we show a reduced synchrony effect for musicians. Taken together, these findings provide evidence for a two-way training-induced mechanism, where musical practice increases sensitivity to audiovisual asynchrony by reducing the brain resources required when multisensory information is obviously synchronous plus fine tuning and increasing precision when a delay is present between the incoming auditory and visual information.

4 Effects of Language Experience on Perception of Audiovisual Synchrony

The vast majority of humans can be regarded as experts in speech perception; however, in general, individuals are only experts in their own native language. Navarra, Alsius [100] took advantage of this fact to explore how expertise with a particular language influences an individual's perception of audiovisual speech synchrony. Synchronous and asynchronous audiovisual stimuli containing either English or Spanish sentences were presented to native English and native Spanish participants, while their task was to decide if the audio and visual streams were in synchrony or not. Native language experience was found to increase the amount of visual lead required for the audio and visual streams to be optimally perceived as synchronous, i.e., when the speech was in the participants' native language their point of subjective simultaneity (PSS) was larger than when it was in the foreign language. Interestingly, this effect was not present in a group of participants who had experience with both English and Spanish [100]. Thus being a speech expert appears to increase participants' tolerance to audiovisual asynchrony, while being a music expert appears to decrease such a tolerance [101]. This posits an interesting question of whether musical and speech expertise have somehow different effects on brain plasticity leading to different PSS when either the visual or the auditory information are unfamiliar (e.g. from a non-native language) to the listener. Here we build on the Navarra, Alsius [100] study and describe a similar experiment that also aimed to explore how mismatches between visual and auditory language (e.g. seeing the facial movements for the native language coupled with the sound of a non-native language) influences the perception of audiovisual speech synchrony.

4.1 Methods

Participants. Eighteen monolingual native English speakers took part in the experiment. All participants had between two and four years of French lessons in school, none had taken Italian, and all described themselves as monolinguals with very little experience of any other language. Nine of the participants were female and the age range was between 17 and 29 (mean = 22).

Stimuli. Stimuli were dynamic audiovisual movies of either, a native English speaker or a native Italian speaker separately saying the words, "tomorrow", "domani" and "andesker". Domani is tomorrow in Italian and andesker is a made up nonsense word. It is worth noting that the English speaker had no experience of speaking Italian, while the native Italian speaker was actually a bilingual Italian/English speaker. All three words have, the same number of syllables (three), a similar spoken duration (about 1 second) and can easily be spoken with neutral affect. Ten cue onset asynchronous (COA) versions of each movie were created: the audio was either shifted to begin before the video (-400, -320, -240, -160, -80ms) or after (+400, +320, +240, +160, +80ms), in 80ms (2 frame) increments. Two versions of each stimulus were created: one containing the full face of the actor the other only the lower half of the face, i.e., the mouth region. In total, there were 132 movies: 2 (speaker - native English or Italian) x 2 (stimulus view - full face or mouth region) x 3 (word - tomorrow, domani, andesker) x 11 (COA levels - ±400, ±320, ±240, ±160, ±80, 0ms).

Procedure. Participants completed three sessions on separate days, each lasting around 45 minutes. During a session there were six experimental blocks, each consisting of one presentation of all 132 movies in random order - giving a total of 18 repetitions of each asynchrony level for each condition. After each movie the task question, "Were the audio and visual streams in synchrony with each other?" and possible answers, "1 for in synch and 3 for out of synch", remained on screen until the participant responded at which point the next trial began.

Analysis first involved, for each subject and each of the twelve conditions, finding the normal Gaussian curve that best fit the number of synchronous responses at each COA level. From this fit to the data two parameters of interest were derived: the point of subjective simultaneity (PSS) and the temporal integration window (TIW). The PSS is derived by taking the millisecond COA value that corresponds to the peak of the best-fitting Gaussian and is generally interpreted as the COA that is perceived as being optimally synchronous [77, 101]. The standard deviation (SD) of the fitted distribution was taken as an estimate of the TIW. This window represents the range of COA, around the PSS, within which participants are unable to reliably perceive asynchrony.

The PSS and TIW data were then used in separate repeated measures 3-factor analysis of variance (ANOVA) tests: 2 (speaker - native English or Italian) x 2 (stimulus view - full face or mouth region) x 3 (word - tomorrow, domani, andesker). It was evident from these tests that there were no significant differences involving the stimulus view condition. Therefore, we collapsed across this condition before refitting each subjects data and re-estimating the PSS and TIW for each of the six remaining conditions: 2 (speaker - native English or Italian) x 3 (word - tomorrow, domani, andesker).

4.2 Results

The 2-factor repeated measures ANOVA on the PSS data highlighted a significant interaction between the factors speaker and word ($F_{2, 34} = 158.87$, $p < 0.001$) and a significant main effect of word ($F_{2, 34} = 8.15$, $p = 0.01$). The main effect of nationality was not significant ($F_{1, 17} = 1.47$, $p = 0.242$). To further explore the interaction,

1-factor ANOVAs on the word factor were run separately for each speaker nationality. Significant main effects for both the native English ($F_{2, 34} = 59.83$, $p < 0.001$) and native Italian speakers ($F_{2, 34} = 21.76$, $p < 0.001$) were found. Bonferroni corrected pairwise comparison follow up tests highlighted the cause of the significant interaction (see also Figure 2): when the native English speaker said "domani" the PSS (126ms) was significantly larger compared to that of "tomorrow" (36ms, $p < 0.001$) or "andesker" (40ms, $p < 0.001$); in contrast, when the native Italian speaker said "domani" the PSS (35ms) was significantly lower compared to "tomorrow" (70ms, $p < 0.001$) and "andesker" (80ms, $p < 0.001$). There was never a significant difference between tomorrow and andesker ($p = 0.753$ for English speaker and $p = .673$ for Italian speaker).

The 2-factor repeated measures ANOVA on the TIW data highlighted a significant interaction between the factors speaker and word ($F_{2, 34} = 13.59$, $p < 0.001$) as well as significant main effects of word ($F_{2, 34} = 4.54$, $p = 0.018$) and nationality ($F_{1, 17} = 5.02$, $p = 0.039$). To further explore the interaction, 1-factor ANOVAs on the word factor were tested separately for each speaker nationality. The ANOVA for the native English speaker produced a significant main effect ($F_{2, 34} = 3.69$, $p = 0.035$); however, bonferroni corrected pairwise comparison follow up tests highlighted no significant difference between any of the words. The ANOVA for the native Italian speaker was also significant ($F_{2, 34} = 7.63$, $p = 0.002$) and follow up tests showed that the TIW for andesker was significantly larger than that of both tomorrow ($p = 0.021$) and domani ($p = 0.036$).

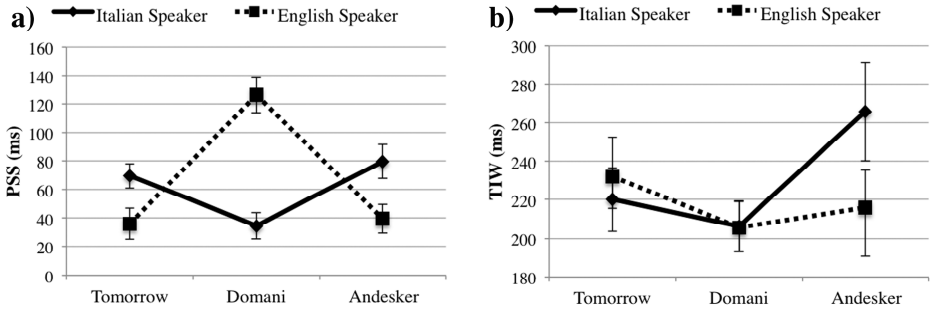


Fig. 2. Average PSS (a) and TIW (b) values for each word spoken by a native English speaker (dashed line with square markers) and a native Italian speaker (solid line with diamond markers). Error bars represent standard errors of mean

4.3 Discussion

To investigate the role of experience with a particular language on synchrony perception we studied synchrony judgments made by native English speakers on audiovisual movies of two speakers (one native English the other native Italian) uttering three different words. Regarding the PSS, a speaker by word interaction was found: while "domani" uttered by the English speaker led to the largest PSS, it led to the smallest PSS when uttered by the Italian speaker. Furthermore, there was no

difference between the English word and the nonsense word regardless of which speaker produced them. Measures of the TIW also highlighted an interaction that was mostly due to the nonsense word producing a particularly large TIW for the Italian speaker compared to the English speaker.

Interestingly, the smallest PSS for each speaker occurred for the speaker's native language. This indicates that rather than the experience of the participants' native language being solely responsible for our results, the experience of the speaker in producing the words greatly influenced synchrony judgments. This is further supported by the fact that the difference in PSS between the two real words (tomorrow and domani) was larger for the monolingual English speaker than the bilingual Italian/English speaker. This suggests that the bilingual speaker produced a more coherent and congruent relationship between the visual and auditory cues to his non-native real word, i.e., tomorrow, than did the monolingual English speaker for his non-native real word, i.e., domani. Therefore, our results demonstrate that while experience is often used to refer to the participants' familiarity with the items being studied, in the particular case of audiovisual speech, the speaker's experience also plays a major role in how their speech will be perceived. This, along with methodological differences, may help to explain why we failed to replicate the results of Navarra, Alsius [100]. Our native English observers required a smaller PSS for native speech compared to non-native speech, which is the opposite result of Navarra et al. Also when the display portraying the Italian speaker (e.g. non-native language movements to the English listeners) was coupled with the native language word 'tomorrow' the PSS was smaller than when the display portraying the English speaker (e.g. native language movements to the listener) was coupled with the non-native word 'domani', meaning that the increased tolerance to audiovisual asynchrony was mostly driven by the non-nativity/unfamiliarity of the sound information.

No differences were observed between the English word and the non-word for both speakers. Experience with the non-words was similar for both speakers in that neither had experience with the 'word' andesker; however, participants reported that andesker sounded like it could be a possible English word. So the comparison between the English word and the non-word demonstrates the role of the participants' experience and expectations on their synchrony judgments. When the speaker's ability is similar, the listeners' expectation/experience is driving their judgments, making the perception of the non-word similar to that of the English word.

Finally, introducing a mismatch between visual and auditory language information causes speech experts to be more tolerant to audiovisual asynchrony (e.g. producing larger PSS) than when the language of the visual and auditory information matches. This result is opposite to what we and others found with music and object actions displays, for which either the mismatch between visual and auditory information reduced the amount of tolerance to audiovisual asynchrony [101] or did not have any effect on temporal discrimination accuracy [102, 103]. Our findings, however, are similar to other speech studies in which gender mismatch between speaker and produced syllables increased how much the visual information had to lead the auditory in order to perceive them simultaneously ([95], see Experiment 1 and 3). Hence, the results of the present study supportVatakis and Spence [102] hypothesis

that the 'unity assumption' (i.e., the observer's assumption that two different sensory signals refer to the same multisensory event) may not have the same effect on the multisensory integration of speech and non-speech stimuli. This supports the idea that the effect of the 'unity assumption' can be driven by both top-down and bottom-up factors contributing to multisensory integration.

5 Summary

In this chapter we discussed the role of experience on human abilities to integrate sight and sound, and related behavioral results to potential neural mechanisms. Moreover, we focused on the topic of synchrony perception in the domains of music and speech expertise. The evidence presented provides clues to some possible mechanisms of multisensory integration that are common across different domains. Clearly our results taken together with the broader literature indicate that experience does alter neural mechanisms in the process of obtaining heightened abilities to discriminate temporal properties. In the future we believe that studies using methods other than those involving the study of temporal synchrony will be helpful to understand differences between these domains. For example, musical training might change the ways in which we weight particular sensory cues and understanding this weighting could provide important insights into the development of expertise.

References

1. Landy, M.S., et al.: Measurement and modeling of depth cue combination: in defense of weak fusion. Vision Res. 35(3), 389–412 (1995)
2. Ernst, M.O., Banks, M.S.: Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415(6870), 429–433 (2002)
3. Green, C.S., Bavelier, D.: Action-video-game experience alters the spatial resolution of vision. Psychol. Sci. 18(1), 88–94 (2007)
4. Simmons, R.W., Locher, P.J.: Role of extended perceptual experience upon haptic perception of nonrepresentational shapes. Percept. Mot. Skills 48(3 Pt. 1), 987–991 (1979)
5. Kisilevsky, B.S., et al.: Effects of experience on fetal voice recognition. Psychol. Sci. 14(3), 220–224 (2003)
6. Atkins, J.E., Fiser, J., Jacobs, R.A.: Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. Vision Res. 41(4), 449–461 (2001)
7. Jacobs, R.A., Fine, I.: Experience-dependent integration of texture and motion cues to depth. Vision Res. 39(24), 4062–4075 (1999)
8. Powers III, A.R., Hillock, A.R., Wallace, M.T.: Perceptual training narrows the temporal window of multisensory binding. J. Neurosci. 29(39), 12265–12674 (2009)
9. Mamassian, P., Goutcher, R.: Prior knowledge on the illumination position. Cognition 81(1), B1-B9 (2001)
10. Mamassian, P., Landy, M.S.: Interaction of visual prior constraints. Vision Res. 41(20), 2653–2668 (2001)
11. Mondloch, C.J., et al.: Face perception during early infancy. Psychol. Sci. 10(5), 419–422 (1999)
12. Turati, C.: Why faces are not special to newborns: An alternative account of the face preference. Current Directions in Psychological Science 13(1), 5–8 (2004)

13. Hershber, W.: Attached-Shadow Orientation Perceived as Depth by Chickens Reared in an Environment Illuminated from Below. Journal of Comparative and Physiological Psychology 73(3), 407-&. (1970)
14. Gregory, R.L.: Knowledge in perception and illusion. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 352(1358), 1121–1127 (1997)
15. Adams, W.J., Graf, E.W., Ernst, M.O.: Experience can change the 'light-from-above' prior. Nat. Neurosci. 7(10), 1057–1058 (2004)
16. Dayan, E., Cohen, L.G.: Neuroplasticity subserving motor skill learning. Neuron 72(3), 443–454 (2011)
17. May, A.: Experience-dependent structural plasticity in the adult human brain. Trends Cogn. Sci. 15(10), 475–482 (2011)
18. Pascual-Leone, A., et al.: The plastic human brain cortex. Annu. Rev. Neurosci. 28, 377–401 (2005)
19. Latinus, M., Crabbe, F., Belin, P.: Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex 21(12), 2820–2828 (2011)
20. Bentin, S., et al.: Electrophysiological Studies of Face Perception in Humans. J. Cogn. Neurosci. 8(6), 551–565 (1996)
21. Rossion, B., et al.: Expertise training with novel objects leads to left-lateralized facelike electrophysiological responses. Psychol. Sci. 13(3), 250–257 (2002)
22. Bukach, C.M., et al.: Does acquisition of Greeble expertise in prosopagnosia rule out a domain-general deficit? Neuropsychologia 50(2), 289–304 (2012)
23. Tanaka, J.W., Curran, T.: A neural basis for expert object recognition. Psychol. Sci. 12(1), 43–47 (2001)
24. Busey, T.A., Vanderkolk, J.R.: Behavioral and electrophysiological evidence for configural processing in fingerprint experts. Vision Res. 45(4), 431–448 (2005)
25. Busey, T.A., Parada, F.J.: The nature of expertise in fingerprint examiners. Psychon. Bull. Rev. 17(2), 155–160 (2010)
26. Calvo-Merino, B., et al.: Action observation and acquired motor skills: an FMRI study with expert dancers. Cereb Cortex 15(8), 1243–1249 (2005)
27. Calvo-Merino, B., et al.: Seeing or doing? Influence of visual and motor familiarity in action observation. Curr. Biol. 16(19), 1905–1910 (2006)
28. Cross, E.S., Hamilton, A.F., Grafton, S.T.: Building a motor simulation de novo: observation of dance by dancers. Neuroimage 31(3), 1257–1267 (2006)
29. Maguire, E.A., et al.: Navigation-related structural change in the hippocampi of taxi drivers. Proc. Natl. Acad. Sci. U S A 97(8), 4398–4403 (2000)
30. Woollett, K., Maguire, E.A.: Acquiring "the Knowledge" of London's layout drives structural brain changes. Curr. Biol. 21(24), 2109–2114 (2011)
31. Hufner, K., et al.: Structural and functional plasticity of the hippocampal formation in professional dancers and slackliners. Hippocampus 21(8), 855–865 (2011)
32. Bezzola, L., et al.: Training-induced neural plasticity in golf novices. J. Neurosci. 31(35), 12444–12448 (2011)
33. Johansen-Berg, H., et al.: Integrity of white matter in the corpus callosum correlates with bimanual co-ordination skills. Neuroimage 36(suppl. 2), T16–T21 (2007)
34. Tuch, D.S., et al.: Choice reaction time performance correlates with diffusion anisotropy in white matter pathways supporting visuospatial attention. Proc. Natl. Acad. Sci. U S A 102(34), 12212–12217 (2005)
35. Draganski, B., et al.: Neuroplasticity: changes in grey matter induced by training. Nature 427(6972), 311–312 (2004)
36. Scholz, J., et al.: Training induces changes in white-matter architecture. Nat. Neurosci. 12(11), 1370–1371 (2009)

37. Bengtsson, S.L., et al.: Extensive piano practicing has regionally specific effects on white matter development. Nat. Neurosci. 8(9), 1148–1150 (2005)
38. Bermudez, P., et al.: Neuroanatomical correlates of musicianship as revealed by cortical thickness and voxel-based morphometry. Cereb Cortex 19(7), 1583–1596 (2009)
39. Gaser, C., Schlaug, G.: Brain structures differ between musicians and non-musicians. J. Neurosci. 23(27), 9240–9245 (2003)
40. Hutchinson, S., et al.: Cerebellar volume of musicians. Cereb Cortex 13(9), 943–949 (2003)
41. Halwani, G.F., et al.: Effects of practice and experience on the arcuate fasciculus: comparing singers, instrumentalists, and non-musicians. Front Psychol. 2, 156 (2011)
42. Imfeld, A., et al.: White matter plasticity in the corticospinal tract of musicians: a diffusion tensor imaging study. Neuroimage 46(3), 600–607 (2009)
43. Schmithorst, V.J., Wilke, M.: Differences in white matter architecture between musicians and non-musicians: a diffusion tensor imaging study. Neurosci. Lett. 321(1-2), 57–60 (2002)
44. Schlaug, G., et al.: In vivo evidence of structural brain asymmetry in musicians. Science 267(5198), 699–701 (1995)
45. Ozturk, A.H., et al.: Morphometric comparison of the human corpus callosum in professional musicians and non-musicians by using in vivo magnetic resonance imaging. J. Neuroradiol. 29(1), 29–34 (2002)
46. Tervaniemi, M.: Musicians–same or different? Ann. N Y Acad. Sci. 1169, 151–156 (2009)
47. Aydin, K., et al.: Quantitative proton MR spectroscopic findings of cortical reorganization in the auditory cortex of musicians. AJNR Am. J. Neuroradiol. 26(1), 128–136 (2005)
48. Elbert, T., et al.: Increased cortical representation of the fingers of the left hand in string players. Science 270(5234), 305–357 (1995)
49. Hyde, K.L., et al.: Musical training shapes structural brain development. J. Neurosci. 29(10), 3019–3025 (2009)
50. Hyde, K.L., et al.: The effects of musical training on structural brain development: a longitudinal study. Ann. N Y Acad. Sci. 1169, 182–186 (2009)
51. Kraus, N., Chandrasekaran, B.: Music training for the development of auditory skills. Nat. Rev. Neurosci. 11(8), 599–605 (2010)
52. Munte, T.F., Altenmuller, E., Jancke, L.: The musician's brain as a model of neuroplasticity. Nat. Rev. Neurosci. 3(6), 473–478 (2002)
53. Bangert, M., Altenmuller, E.O.: Mapping perception to action in piano practice: a longitudinal DC-EEG study. BMC Neurosci. 4, 26 (2003)
54. Magne, C., Schon, D., Besson, M.: Musician children detect pitch violations in both music and language better than nonmusician children: behavioral and electrophysiological approaches. J. Cogn. Neurosci. 18(2), 199–211 (2006)
55. Milovanov, R., Tervaniemi, M.: The Interplay between Musical and Linguistic Aptitudes: A Review. Front Psychol. 2, 321 (2011)
56. Chan, A.S., Ho, Y.C., Cheung, M.C.: Music training improves verbal memory. Nature 396(6707), 128 (1998)
57. Aleman, A., et al.: Music training and mental imagery ability. Neuropsychologia 38(12), 1664–1668 (2000)
58. Brochard, R., Dufour, A., Despres, O.: Effect of musical expertise on visuospatial abilities: evidence from reaction times and mental imagery. Brain Cogn. 54(2), 103–109 (2004)

59. Schmithorst, V.J., Holland, S.K.: The effect of musical training on the neural correlates of math processing: a functional magnetic resonance imaging study in humans. Neurosci. Lett. 354(3), 193–196 (2004)
60. Besson, M., Chobert, J., Marie, C.: Transfer of Training between Music and Speech: Common Processing, Attention, and Memory. Front Psychol. 2, 94 (2011)
61. Patel, A.D.: Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. Front Psychol. 2, 142 (2011)
62. Wallace, M.T., Stein, B.E.: Sensory and multisensory responses in the newborn monkey superior colliculus. J. Neurosci. 21(22), 8886–8894 (2001)
63. Wallace, M.T., Stein, B.E.: Development of multisensory neurons and multisensory integration in cat superior colliculus. J. Neurosci. 17(7), 2429–2444 (1997)
64. Wallace, M.T., Stein, B.E.: Cross-modal synthesis in the midbrain depends on input from cortex. J. Neurophysiol. 71(1), 429–4232 (1994)
65. Jiang, W., Jiang, H., Stein, B.E.: Neonatal cortical ablation disrupts multisensory development in superior colliculus. J. Neurophysiol. 95(3), 1380–1396 (2006)
66. Haslinger, B., et al.: Transmodal sensorimotor networks during action observation in professional pianists. J. Cogn. Neurosci. 17(2), 282–293 (2005)
67. Hodges, D.A., Hairston, W.D., Burdette, J.H.: Aspects of multisensory perception: the integration of visual and auditory information in musical experiences. Ann. N Y Acad. Sci. 1060, 175–185 (2005)
68. Petrini, K., et al.: Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. Exp. Brain Res. 198(2-3), 339–352 (2009)
69. Petrini, K., Russell, M., Pollick, F.: When knowing can replace seeing in audiovisual integration of actions. Cognition 110(3), 432–439 (2009)
70. Petrini, K., Holt, S.P., Pollick, F.: Expertise with multisensory events eliminates the effect of biological motion rotation on audiovisual synchrony perception. J. Vis. 10(5), 2 (2010)
71. Petrini, K., et al.: Action expertise reduces brain activity for audiovisual matching actions: an fMRI study with expert drummers. Neuroimage 56, 1480–1492 (2011)
72. Musacchia, G., et al.: Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. Proc. Natl. Acad. Sci. U S A 104(40), 15894–15898 (2007)
73. Schulz, M., Ross, B., Pantev, C.: Evidence for training-induced crossmodal reorganization of cortical functions in trumpet players. Neuroreport 14(1), 157–161 (2003)
74. Lappe, C., et al.: Cortical plasticity induced by short-term unimodal and multimodal musical training. J. Neurosci. 28(39), 9632–9639 (2008)
75. Schlaug, G., et al.: Effects of music training on the child's brain and cognitive development. Ann. N Y Acad. Sci. 1060, 219–230 (2005)
76. Spence, C., Squire, S.: Multisensory integration: maintaining the perception of synchrony. Curr. Biol. 13(13), R519–R521 (2003)
77. Arrighi, R., Alais, D., Burr, D.: Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. J. Vis. 6(3), 260–268 (2006)
78. van Wassenhove, V., Grant, K.W., Poeppel, D.: Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45(3), 598–607 (2007)
79. Vatakis, A., Spence, C.: Audiovisual synchrony perception for music, speech, and object actions. Brain Res. 1111(1), 134–142 (2006)
80. Vatakis, A., Spence, C.: Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. Neurosci. Lett. 393(1), 40–44 (2006)

81. Dixon, N.F., Spitz, L.: The detection of auditory visual desynchrony. Perception 9(6), 719–721 (1980)
82. Lee, H., Noppeney, U.: Long-term music training tunes how the brain temporally binds signals from multiple senses. Proc. Natl. Acad. Sci. U S A 108(51), E1441–E1450 (2011)
83. Petrini, K., et al.: The music of your emotions: neural substrates involved in detection of emotional correspondence between auditory and visual music actions. PLoS One 6(4), e19165 (2011)
84. Petrini, K., McAleer, P., Pollick, F.: Audiovisual integration of emotional signals from music improvisation does not depend on temporal correspondence. Brain Res. 1323, 139–148 (2010)
85. Hein, G., et al.: Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. J. Neurosci. 27(30), 7881–7887 (2007)
86. Kim, R.S., Seitz, A.R., Shams, L.: Benefits of stimulus congruency for multisensory facilitation of visual learning. PLoS One 3(1), e1532 (2008)
87. Munhall, K.G., et al.: Temporal constraints on the McGurk effect. Percept. Psychophys 58(3), 351–362 (1996)
88. Vatakis, A., et al.: Temporal recalibration during asynchronous audiovisual speech perception. Exp. Brain Res. 181(1), 173–181 (2007)
89. Fain, G.L.: Sensory transduction, 340 p. Sinauer Associates, Sunderland (2003)
90. King, A.J.: Multisensory integration: strategies for synchronization. Curr. Biol. 15(9), R339–R3941 (2005)
91. King, A.J., Palmer, A.R.: Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. Exp. Brain Res. 60(3), 492–500 (1985)
92. Waadeland, C.H.: Strategies in empirical studies of swing groove. Musicologia Norvegica 32, 169–191 (2006)
93. Jansson, G., Johansson, G.: Visual perception of bending motion. Perception 2(3), 321–326 (1973)
94. McGurk, H.M., Macdonald, J.: Hearing lips and seeing voices. Nature 264(5588), 746–748 (1976)
95. Vatakis, A., Spence, C.: Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Percept. Psychophys 69(5), 744–756 (2007)
96. Drake, C., Jones, M.R., Baruch, C.: The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending. Cognition 77(3), 251–288 (2000)
97. Saygin, A.P., Driver, J., de Sa, V.R.: In the footsteps of biological motion and multisensory perception: judgments of audiovisual temporal relations are enhanced for upright walkers. Psychol. Sci. 19(5), 469–475 (2008)
98. Aschersleben, G., Prinz, W.: Synchronizing actions with events: the role of sensory information. Percept. Psychophys 57(3), 305–317 (1995)
99. Miyake, Y., Onishi, Y., Poppel, E.: Two types of anticipation in synchronization tapping. Acta Neurobiol Exp. (Wars), 64(3), 415–426 (2004)
100. Navarra, J., et al.: Perception of audiovisual speech synchrony for native and non-native language. Brain Res. 1323, 84–93 (2010)
101. Petrini, K., et al.: Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. Experimental Brain Research 198(2-3), 339–352 (2009)
102. Vatakis, A., Spence, C.: Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. Acta Psychol. (Amst) 127(1), 12–23 (2008)
103. Vatakis, A., Ghazanfar, A.A., Spence, C.: Facilitation of multisensory integration by the "unity effect" reveals that speech is special. J. Vis. 8(9),14. 1–14.11 (2008)

Nonverbal Communication – Signals, Conventions and Incommensurable Explanations

Lutz-Michael Alisch

Technische Universität Dresden,
Faculty of Educational Research, Institute 1,
D-1062 Dresden, Germany
ilona.lutz.alisch@t-online.de

Abstract. Considering nonverbal communication and its complexity, four problems are addressed which focus on the dynamics of nonverbal communication. (1) How much of the complexity of nonverbal communication is due to the amount of expressions following cultural rules (e. g. conventions) and the expressions of the agents´ states through their signaling systems? (2) Nonverbal behaviour can be regarded as time-varying multi-scaled multimodal configurations of magnitudes. It is natural to ask for scaling laws. (3) Furthermore, the dynamics of the configurations just mentioned is of interest. (4) Why are verbal expressions more conventional than nonverbal ones? The discussion of the four problems suggests that signal-based explanations of nonverbal behaviour and communication are incommensurable with convention-based explanations.

Keywords: Signaling systems, multi-scaled multimodal configurations, configuration dynamics, nonverbal conventions.

1 Introduction

The excellent contributions made by so many scientists to the COST Action 2102 are impressive. Many of them present a fascinating exploration of how fundamental nonverbal signals are to communication and how intelligible engineering solutions have to be to enhance the communicative abilities of both humans and human-like machines. It seems to me, that the following quotation from Brian Skyrms, a logician and philosopher of science, could be regarded as a kind of reflective résumé. Information flows throughout communication.

> "But there is a lot more going on than the simple transmission of information. Information is filtered, combined, and processed … Signals effect coordinated behaviour between multiple actors – often for mutual benefit. The details of how this works in even the simplest natural signaling networks are amazingly complex" [40], p. 177.

It is this complexity that represents the core of our common scientific challenges notwithstanding the efforts and the multiplicity of successes which must be acknowledged.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 321–334, 2012.

In what follows I would like to indicate how we can bring fundamental theoretical tools to bear on some problems which are intimately interweaved with nonverbal communication and its complexity. Whatever one thinks of fundamental theoretical tools, a philosophical flavor could not be excluded. I propose to address myself to four problems. My interest in these problems is certainly dependent on my own view of nonverbal communication and on my personal interests in modeling and synthesizing even seemingly heterogeneous results. My approach also makes use of some philosophy of science. I have some doubts on the universality of the sense-think-express cycle when explaining or constructing systems which can evoke nonverbal behaviour. In general, I am worrying about cases in which a lack of understanding nonverbal communication is manifested. I do not present a full-fledged theory but only a part of a strategy of normal science which means muddling through quite nicely.

The four problems just mentioned focus upon the dynamics of nonverbal communication, more specifically: (1) Beyond the scope of phenomenology the complexity of nonverbal communication corresponds to the dynamics of the communicating agents and to the static and dynamic variables of the environment in which the agents behave. It is helpful to distinguish signals from cultural rules like conventions and meaning, but the difficulty remains determining how much of the complexity is due to the dynamics of the agents and how much is due to their environment. Furthermore, we are not able until now to determine the amount of nonverbal expressions due to following conventions and the amount due to the expressions of the agents´ states through their signaling systems.

(2) Nonverbal behaviour is multimodal. However, this is an insufficient description. Each modality consists of orders of magnitudes. The entire variables and parameters take their values in time and the values are related to form a time-dependent configuration. It is natural to ask for scaling laws and furthermore for the relevance of the constructal law [8] in the context of nonverbal communication.

(3) The time-varying multi-scaled multimodal configurations of nonverbal magnitudes differ with respect to the dynamics inherent in their order and their spatio-temporal evolvement. Apart from questions with respect to these two sorts of nonverbal dynamics it is interesting to ask for the role of deterministic and probabilistic aspects in nonverbal communication.

(4) Our verbal languages are instructive examples of highly regulated canonical systems of conventions. However this does not hold exactly for nonverbal expressions in everyday communication. Of course, an analysis of performing arts would indicate a higher level of regulated nonverbal behaviour. Compensatory sign languages are even invented to be available in biprojective correspondence to larger parts of verbal languages. Nevertheless, verbal languages are more conventional than nonverbal expressions. Why this is the case? From an evolutionary point of view, nonverbal communication has had much more time to change into a system of conventions than verbal languages [44]. Why didn't this happen? This is my fourth problem.

The discussion of the four problems suggests that there is a gap between signals and conventions. As a consequence signal-based explanations of nonverbal communication and convention-based explanations are incommensurable, at least

with respect to the present state-of-the-art. Before getting into details, a note on incommensurability is given (thanks to Vincent Müller and Anna Esposito who insisted on a clarification). Due to the many facets of incommensurability [47] it is necessary to give a comment upon the version which is used here. Both, Kuhn [24] and Feyerabend [16] are said to maintain that rivaling theories use conceptual schemes which are not comparable or even untranslatable [21], chap. 5. Synonomy in the meaning of scientific terms respectively is denied. A somewhat weaker version of incommensurability claims that the concepts of the theories are not coreferential [17], [36]. It has been suggested that a useful distinction should be made between the concepts of a theory and its consequences [39]. With this is mind, the untranslatability of the concepts of the two theories is trivial in a certain sense. A language which is not universal but serves the purpose to cover a restricted terrain could not be translated into a language with another area of reference. Any case in which nonuniversal languages are untranslatable in a nontrivial way is a case of incompatibility. Note that in such a case the missing semantic harmony between the concepts of the two theories does not prevent the reduction of the theories [39].

Considering the consequences of theories, Kuhn [24] and Feyerabend [16] claim that untranslatability is followed by an incomparability of the consequences of the two theories. This seems to be a correct statement but only in case of logical deduction of a consequence of one of the theories in terms of the other. As soon as inferential deduction is replaced by inferential asymptotics in the context of approximate reduction, the incommensurability of the consequences of two theories can be denied.

Perhaps with the Kuhnian [25] analogy in mind between the mathematical statement “no common measure” and the empirical statement “no common language”, some authors have assumed that incommensurability means incompatibility of measurements, which could no longer be regarded as a local semantic defect. Differences in measures come up as differences in observables, a kind of untransformability which indicates likewise semantic untranslatability and furthermore that incommensurability differs from a logical contradiction. A contradiction presuposes a kind of companionship of the two theories in following the rules of the same language whereas incommensurability presupposes two languages which are untranslatable.

The version of incommensurability adopted in what follows is based on a shift from a certain problem (incomparability of two things) to its reverse (if two things are not part of the same problem, then they are incommensurable). Below, it will be shown, that signals and conventions differ in their scale-order. Hence, every kind of mutual reducibility or indentifiability seem to be excluded. However, the existence of different scale-orders do not imply per se unrelatedness. The scales can be connected (e. g. concurrency of the dynamics realized in the scales, bottom-up induction of a state in the higher-order scale through accumulation of states in the lower-order scale and thresholding of the magnitude of the accumulation; top-down reduction; higher-order states emerging out of lower-order states). The connection between scales is a kind of an empirical law and the problem with respect to multi-scaled systems is how could this law be indentified. The problem is called a multi-scaling problem and its

solution is a functional characterization of the multi-scale structure and the dynamics realized as a mapping of the structure inter itself. Now, if two distinct scale-orders (including variables, parameters, magnitudes) are part of a common multi-scaling problem then they are commensurable. If they are not they are incommensurable.

2 Signals, Conventions and Strategies of Measurement

In my own discipline I am often engaged in educational research projects concerning the competencies of teachers. It is well known that teachers' nonverbal immediacy behaviour has impact on affective learning. When both verbal and nonverbal behaviour is considered, the nonverbal behaviour has more influence [32]. The concept of immediacy is defined as perceived closeness between people. Immediacy behaviours are those that result in people perceiving others to be close or distant interactionally [32], p. 423. Immediacy behaviours are multichannelled [4]. The channels are discrete yet interdependent [3] which means that nonverbal cues are processed as a gestalt typically conveyed through proxemic, haptic, oculesic, kinesic, vocalic, and chronemic behaviour simultaneously. The dimensions which characterize immediacy behaviours or, to be more precise, the dimensions of the relevant state space cover involvement, gesture, movement, posture and prosodic variety. Specifically the subdimensions of involvement are differentiated into eye-contact, interpersonal distance, head movement, body orientation and mimic expression [5].

To study nonverbal immediacy in detail, several authors created different measures. One way to measure immediate behaviour is based on self-reports. Another way is to code each videotaped nonverbal behaviour. To compensate some disadvantages of coding such as the inordinately time-consume to code multiple channels of nonverbal behaviour or the missings of behaviour meaningful for an actual interaction, it has been suggested to use supplementary measurement devices such as scales which measure an interactant´s perception of a partner´s immediacy or which give global information about general immediacy or which are used by trained observers to assess immediacy [3], p 116.

All these measures of immediate nonverbal behaviour take into account what may be called conventions. Observers rate immediacy using verbal categories. Teachers or agents answer questions concerning the details of their self-reports. The individuation of nonverbal behaviour channels and the identification of nonverbal behavioural acts are intimately related to the perceptual abilities of the observers and related to their understanding of the meaning of the acts. However, many authors have demonstrated, that convention-based high inference research is biased. One of the most prominent sources of what Klaus Beck [7] has called the semantic bias is the variety of the meaning of both verbal and nonverbal signals. For example, Brugman [10] had studied the relationships among the senses of a single lexical item. She considered nearly 100 senses of the item "over". Lindner [31] took up the question: How are the senses of particles (such as out of figure out, space out, fill out) related to one another? She reported more than 600 examples of out and more than 1200 examples of up in verb-particle construction (for further details see [26]).

Most of the measures of immediate nonverbal behaviour are constructed as performance assessment scoring systems. Raters look for evidence of particular aspects of immediate behaviour and place that immediacy on a scale [35]. In another way, the raters give response on a stimulus presented via an item. For each response, a score within a whole scoring system is defined. As just mentioned, Beck has shown that meaning differences and semantic variability between the responses of the raters result in sometimes even dramatic scoring differences.

It is natural to look for improvements to get scientifically better measures. The strategy obtained is called psychologizing the conventions. In a first step logico-semantic analysis conducts to lay open the elementary components of a convention. Then it is convenient to find psychological laws which connect the components with signal-based regulators. Epistemologically the strategy of psychologizing is realistic in the sense of establishing a presumption and then validating the existence of the designated reference. In nonverbal communication research realism means to set off with a qualitative description of a phenomenologically individuated convention and to proceed with an anchoring of the conventional meaning in patterned signals. In short, progress in measuring immediate nonverbal behaviour consists of a transition from high-inference to low-inference research.

With the denotation of naturalizing nonverbal communication research, the measurement strategies are in a certain sense in reverse to the realistic measures. Naturalized measurement is the very part of low-inference research. The objects that are measured can be described as signaling systems and the data captured by non-human measurement devices are signals. Epistemologically, naturalizing is an empiricistic strategy. Here, empiricism means to set off with signals as data and then to proceed with pattern recognition and specifically with the search for correlations to meaningful nonverbal expressions and conventions.

In terms of the theory of measurement, psychologizing is representational. A qualitative structure of a measurable attribute given as a set of qualitative axioms (a set-theoretic predicate) is represented through a homomorphism by a numeric structure, given as a set of theoretic axioms. Often, Stevens [43] is quoted for an abbreviation of a definition of measurement as the assignment of numbers to attributes according to a rule. In contrast, measurement in the context of naturalizing follows the interactionistic approach, which goes back to the classical distinction between intensive and extensive quantities. Measurement is based on a physical interaction between an instrument by which the measurement is carried out and a system on which the measurement is being done [13], p. 198. To get the desired type of quantity at the end of the interaction, one has to bring together the dual pairs of spaces of intensive and extensive quantities by an inner product (defined by the usual expectation integral). For example, Radon-Nikodym derivatives produce intensive quantities from pairs of extensive quantities [13], p. 199. According to the interactionistic view, measurement is the production of the numerical value of some empirically realized inner product.

In the theory of measurement, it has been suggested that there must be a framework for reasoning about the representationalistic and the interactionistic view in order to fulfill the desire of unification. I do agree with this suggestion. However,

as far as I know, the framework is rather a sketch than a unified theory. As a consequence, data obtained by a representational strategy differ not only mutually dependent on their qualitative axiomatic characterization but also from data obtained by some empirically realized inner product. Note, that this difference is not on principle. Note also, that at the moment we do not have an ensured measure theoretic relation between conventions and signals.

What do I mean, when I am talking about the concepts of signals and conventions? Often, signals are defined as functions of one or several variables transmitting information with respect to the state of an observed system. The information is transmitted by signaling systems at all levels of biological organization [40], p. 6. Some of these signaling systems are innate, some are not. In general, signals are not endowed with any intrinsic meaning.

In the year 1754, Jean-Jacques Rousseau had asked for an explanation of the origins of meaning without the use of a meaningful language. This is it, what linguists call the problem of the ultra-history of the origins of meaning. Paul Grice [20] distinguished between natural and non-natural meaning. Brian Skyrms translated this into the language of signals while pointing to Grice´s distinction as the distinction between conventional and non-conventional meaning. He insists that conventional meaning is a variety of natural meaning. According to Skyrms, evolution and learning as natural dynamic processes create conventions.

The problem stated by Rousseau is perhaps not solvable in general. However, David Lewis [28], [29] used signaling games to prove the existence of a singular solution to Rousseau´s problem. This solution could be regarded as a lower bound for the space of all possible solutions. Indeed, a convention can be realized without pre-existing conventions. It really works, but we do not know if this is a sufficient explanation of the creation of meaningful nonverbal behaviour from signals. Lewis defined conventions as the solutions of some game theoretic coordination problems. A coordination problem is a situation of interdependent decision by two or more agents in which coincidence of interest predominates and in which there are two or more proper coordination equilibria. A coordination problem is a situation of interdependent decision by two or more agents in which coincidence of interest predominates and in which there are two or more proper coordination equilibria [30], p. 24. When the transmission of information between the players of the game is perfect, so that the act always matches the state and the payoff is optimal, an equilibrium is called a signaling system [40], p. 7. If we switch the messages around the same way in all players, we get the same payoffs. Permutation of messages takes an equilibrium into another. This symmetry is what makes signaling games a model in which the meaning of signals is conventional. Lewis [29] defined conventions more explicitly: X is a convention if in population P and recurrent situation S there exists a regularity R in the behaviour of members of P who are agents in S such that

(1) almost every member of P conforms to R in S;
(2) almost every member of P is convinced to that almost every other member of P conforms to R in S; (3) on account of (2) almost every member of P has a deontic preferable reason to conform to R in S; (4) almost every member of P has the

same preference for "Every member of P conforms to R in S" over "Nearly every member of P conforms to R in S";
(5) there is R´≠ R and R´, R are incommensurable and R´ fulfills (1) and (2). This condition ensures that conventions could be arbitrary;
(6) almost every member of P knows that conditions (1) – (5) are fulfilled by R.

3 Multiscaling

According to the definitions of the concepts of signals and conventions, the strategy of psychologizing is concerned with the identification of signaling games as the basis of every convention in nonverbal behaviour. In contrast to this, the strategy of naturalizing is concerned with the identification of patterned signal sequences. One has to demonstrate that the sequences are part of a signaling game. As long as the gap between signals and conventions exists the explanations of nonverbal behaviour and nonverbal communication based on signals or based on conventions are incommensurable.

I will attempt to illustrate this in what follows. First, I would like to ask how much of a person´s nonverbal behaviour is the result of his/her system dynamics, and how much is the result of both static and dynamic environmental variables [12], p. 3. It seems that we have a tendency to underestimate the dynamics and to overestimate the environment. Conventions are environmental variables which function as control parameters of the nonverbal dynamics. Control parameters open up the history of nonverbal behaviour but do not cause the succession of its states. In this sense, conventions are not dynamic causes of nonverbal behaviour but dynamic noise. Many researchers suggest that we could understand the nonverbal behaviour of a partner when it comes to our conscious minds. In educational research it is expected that about 70 percent of the relevant information in instructional communication is expressed nonverbally and brain researchers tell us that about 70 percent of these nonverbal expressions are processed unconsciously. Whether or not this is exactly the truth, we have to comprise that a remarkable amount of nonverbally transmitted information does not come up in the format of conventions. Furthermore the sense-think-express cycle seems to tell us not the whole story. Eventually, it has a cousin like a sense-process-transit cycle where process is due to signal processing not necessary done consciously and transit is due to the action of the dynamics which drives the time- or space-dependent transitions of the states of the behavioural system. These states are configurations in character which means that an embodied system has a finite set of possible multi-scaled state-parameter values that could be combined under some rules. Suppose a person is in a highly emotional state. Could the person realize the state without emitting nonverbal signals? Moreover, are these signals expressions of the emotional state? In the configuration view it is stated that humans as behavioural systems are at least in principle describable. The state space used for a description represents all the dimensions which are necessary to give the complete

state of the system. The space is structured in that the dimensions are related. The structure which is relevant for an analysis of configurations is a multi-scaled structure. The system cannot suspend any dimension in any scale when it realizes a state. However, the dimensional parameters could have time-varying values and the multi-scale order also could be variable. Every dimension in every scale is ever present but of course not in the same degree at any time. In a multi-scale analysis this results in changeable orders over time. Thus, we have two important properties of configurations. A configuration is composed by the parameter values of all dimensions in all scale orders of the state space of the system. The multi-scaled structure of the configuration is time-dependent due to the time-varying parameter values.

In our example of a system being in a highly emotional state the nonverbal signals are expressions of the emotion only in case of an additional cross-scale relation, for example an intentional dynamic which interacts with the emotion and the emission of the signals. Yet, not every nonverbal companion of an emotional state is an expression of the state. Here, the difficulty of the issues of recognition is being settled. Nonverbal behaviour as part of a state configuration is tied to other parts or scales by cross-scale dynamics. This is what has been called a multi-scaling problem. Solutions of such a problem consist of an identification of pertinent scaling laws (top-down, bottom-up, concurrent etc.).

Recently it has been suggested by Bejan [8], [9] what is now called the constructal law of the generation of flow configuration: For a finite size flow system to persist in time its configuration must evolve in such a way that it provides easier and easier access to the currents that flow through it.

Disregard the dependence of the law on the criterion for “easier”. It could easily be substituted by a formulation which concerns the approximation of an attractor by self-organized attractor hopping. It is more interesting that the constructal law could make available a bridge between the problem concerning the evolvement of conventions from signals on the one hand and the multi-scale dynamics on the other. A first step for an exploration is in the line of naturalizing nonverbal communication research. Let me give an illustration from one of my own research projects. Previously, it is necessary to mention that the preliminarity of my approach is beyond question. It is part of what could be called a hybrid of empirical and synthetic nonverbal research. As a paradigmatic piece of work in purely synthetic nonverbal research I would characterize the computational and robotic synthesis of language evolution which Luc Steels [41], [42] has put forward. From this research I borrow the curiosity about what configurations can tell us. On the other hand the set of results concerning nonverbal behaviour obtained by the empirical nonverbal communication research should give me a clue to relevant phenomena. The wedding of these two sources of inspiration could be celebrated in the context of multi-scaling problems. If my presumption is not totally wrong the interplay between the configuration dynamics and conventions should fit into the constructal law. However, to this end I will have to go still a longer way. That leads over to the third of my four problems.

4 Multimodal Multi-scaled Configuration Dynamics

The third problem is connected with the question: How much nonverbally relevant information is in a sequence of signals? A nonverbal act consists of a configuration of modalities. To be more precise, each modality can be described as an averaged curve in a suitable multidimensional state space. The coordinates of the points of this curve are given by the values of their state parameters which correspond to the dimensions. Hence, a nonverbal act could be defined as a dynamic configuration of multimodal parameter values. The values of the parameters are time-varying. Moreover, the configuration is not only parameterized in time but also in space. Suppose, the distance between a finger tip and the forehead changes in time, so it also changes in space. According to this, nonverbal acts should be defined as spatio-temporal configurations of multimodal parameter values. Another important issue is the scale order of the parameters, which results from different importance of the modalities and their state parameters in different nonverbal acts. Here, importance is not a deontic term but an outcome of a multi-scale analysis. The scale order of a configuration is also time-dependent. My final definition is therefore: A nonverbal act is a spatio-temporal configuration of multi-scaled multimodal parameter values.

It is easy to distinguish between the local and global dynamics of configurations. For example, immediate nonverbal behaviour designates a quality of global configuration dynamics. On the other hand, if the configuration dynamic is stationary in a time interval then we say that the dynamic is local. In one of my research projects we study immediacy behaviours of teachers. Our guess is, that the behaviour of teachers is the more immediate the better the global nonverbal configuration dynamic. However, it seems not quite clear what does better mean in this context. The specification of the term requires a concise analysis of the dynamics.

Our research is settled at the strategy of naturalizing, but with a bow to the convention-based research on immediacy behaviour. The applied measuring instrument collects signals which correspond to some of the state space dimensions mentioned above (gesture, movement, interpersonal distance, head movement, body orientation, spatial behaviour). The instrument is a 3-D camera which simultaneously provides intensity and range images. Methods of range image sequence processing and analysis are developed which allow for 3-D tracking of body points in time (for details see [48], [22]).

Other data are collected with 2D-video cameras to identify different student perspectives on the behaviour of the teacher and for the purpose to code different events in the course of a lesson (for example opening, motivation, instructional talk, class management). Furthermore, the immediacy behaviour of the teacher is measured by student rating with NIS-0, the nonverbal immediacy scale. Testing the 3D-data of ergodicity [14] gives a hint for data analysis. If the dynamics underlying the measured and tracked points is ergodic, the curves of each point of each teacher could be regarded as members of samples accessible for functional data analysis. Otherwise, the trajectories diverge.

It is important to look for nonverbal modalities. We use random spatial sampling [33] to distribute 50 points per each teacher on the immediacy dimensions under

investigation. Tracking and smoothing these points produce 50 curves per each videographed time interval. The curves of each dimension are registrated [37] and then averaged to get a mean function which can be defined as a nonverbal modality. For each teacher in each modality, we estimate baselines and variances.

With respect to immediacy behaviours three different explanations have been suggested. The functional factor analytic model states that immediate behaviour depends on some functional properties of the dynamic of the modalities over several time intervals. The approach based on dynamic matrices claims that immediate behaviour is linked to general characteristics of the nonverbal expression in each modality. The dynamic factor model holds that immediate behaviour traces back to the professional competencies of the teacher. The three models explain immediate behaviour respectively as time dependent or with respect to a structure or controlled by some spatio-temporal influences.

We select just the model which predicts the outcomes of the ratings of the immediacy behaviours best from the modality curves. For this purpose, we use a nonparametric functional linear predictive approach [23]. While the three models take latent variables into consideration, the nonverbal configurations and their spatio-temporal variation are modelled directly. If we could find a relation between the selected one of the three models and the configuration dynamics, then we could get a solution for a multi-scaling problem. A way of modelling the configuration dynamics is the multi-scale analysis of complex time-series of the stationary process in each respective time interval [19]; for a test for stationarity [46]. On account of the stationarity the time-series could be expressed in form of local stationary wavelets . Hence, it is feasible to determine the variances from the wavelet-scales and the localizations in time. Now, the variances can be used to estimate the change points and in a first approximation the global configuration dynamics could be modelled as sequencing the change points.

It was mentioned that this research project is quite preliminary in that the multi-scaling problem seems to be confounded with the signal vs. convention problem. However, if there is something nonverbally relevant in the patterns of the signal sequences, we should notice it at least on the level of the global configuration dynamics. To give a better illustration, I will refer to a second research project. Before, let me remark something concerning determinism and probabilism in nonverbal communication research. The analysis of data from the immediacy project indicates that the underlying dynamics are purely deterministic. Sources of randomness come from the time-dependent variability of the coefficients. Therefore, we get a specific subset of random dynamical systems, namely, random differential equations [1], to describe nonverbal immediacy behaviour. Here, randomness is due to the environment of the agent. Notice, that an intention to express an emotion, for example, belongs to the environment of the observed dynamics.

Another source of randomness is due to the perception of an agent, who communicates with several other persons. In this case, the things are quite different. My second research project is pursued together with Rüdiger Hoffmann. It is concerned with prosodic expressions of teachers in the opening of a lesson. When the teacher enters a classroom, at least in Germany, the students talk together, amuse

themselves, make a lot of noise, perhaps even romp, in short they are very active, vivacious and lively. In order to begin the instruction, the teacher has to appease the students. The technical term in German calls 'einstimmen' which is meant in the sense of 'to get into the mood for'. The teacher tries to stimulate the students to get into a mood for the lesson. In this situation, the proper classroom management technique of teachers with high competencies is to use suitable prosodic expressions. Our guess is that such expressions are successful, when certain sequences of prosodic configurations are realized. I will not go into the details of the research project here (for these details see [2]) but refer to an interesting aspect. When the teacher enters the classroom, he/she perceives the verbal expressions of the students not as singular communicative acts but as noisy occurrences. Hence, the prosodic stimulation to get into a desired mood in essence is an attempt to control the noise. In our research project, we are trying to estimate the noise and synchronize it with prosodic configurations emitted by the teacher.

Unfortunately, the noise is not as domesticated as Gaussian noise, for example. In particular, the trajectories of the prosody of the students are corrupted by impulsive noise [38]. This is challenging. It is customary to model noisy processes with SDE (stochastic differential equations). However, impulsive noise does not correspond to a Wiener process. In technical terms, the process of the prosodic expressions of the students can be considered as a stochastic process with rough paths. Yet, the methods useful for estimating equations with rough paths are still rare [34]. Moreover, the Itō calculus is for no use to get analytic solutions [18]. Furthermore, there is another complication. We are interested in patterns of prosodic teacher expressions which have an effect on the multidimensional stochastic process that describes the prosodic expressions of the students. The teachers' prosodic signals can be transformed in a multiple vector-valued time-series. Multidimensional minimizing splines for spatio-temporal smoothing yield functional prosodic data. Now, we have two processes, the rough paths process of the students and the prosodic expression of the teacher. In order to establish results concerning the mutual influences between the two processes we are looking for synchronizing and desynchronizing patterns. However, this requires the comparison of two stochastic processes. Fajardo and Keisler suggest a method which uses model theory, non-standard analysis and probability theory. Their approach seems to be theoretically elegant, but not directly applicable to data.

To sum up, despite of the technical difficulties just demonstrated, we think of certain patterns of multi-scaled prosodic configurations which could better appease the students via prosodic synchronizing and desynchronizing the two processes just mentioned. We look at this as an example for patterns of configurations that work well though they are not conventions and presumably could not fully be processed in a conscious mode. We are convinced of some reward of our approach, because its results could give evidence not only for classroom management efficacy but also for man-machine communication in order to make the generation of nonverbal behaviour more effective for embodied agents enhancing existing approaches [27]. And last but not least, we feel that this area of nonverbal research presents another example of incommensurable explanations of nonverbal communication, where signals do not correspond to conventions directly.

5 The Fourth Problem

My last problem is the deepest of the four problems. If I trust the statements of Clark and Lappin [11], p. 2, we do not yet know anything substantive about how learning products are represented in the brain whether they are encoded as propositions, or are emergent properties of neural network actions. We do not even have the evidence necessary to formulate the pertinent questions precisely. Instead, there is hardly any doubt, that the evolvement and greater amount of verbal conventions in contrast to nonverbal ones has something to do with the organization of the brain. Let me introduce a highly speculative idea. In several papers, Guilio Tononi has suggested that consciousness depends on integrated information defined as the amount of information generated by a system in a given state, above and beyond the information generated independently by its parts [6], p. 2. Tononi has computed the quantity of integrated information and described the structure or quality of the integrated information unfolded by interactions in the system. The quantity of consciousness corresponds to the amount of integrated information and the quality of conscious experience is specified by the informational relationships [45], p. 216.

Let us recollect my guess that dynamic patterns of configurations hold nonverbal information relevant to communication but not necessarily processed in a conscious mode. Maybe, these patterns do not form a basis to create a convention. Why is that the case? My speculative answer goes like this. Their degree of multimodal multi-scaled information integration and relatedness is not high enough to cause (in the sense of Dretske [15]) a conscious experience. If my idea is not totally wrong, it could shed some light on the role played by the constructal law. Thanks to it, dynamic patterns work well in nonverbal communication though not as conscious conventions. Does this mean that signal-based and convention-based explanations of nonverbal communication are incommensurable in principle? It must not be the case if the basis to form conventions consists of a dynamic signal pattern as a whole being independent of the configurations as its parts. At least it should be clear that for a full understanding of nonverbal communication some information about what is called cognizing [11], p. 2, should be available which means knowledge that is not conscious and is not epistemicly justified. However, this is another story which may be entitled cognitive behavioural systems.

References

1. Alisch, L.-M., Robitzsch, A.: Estimating Dynamical Systems Disturbed by Noise. In: van Dijkum, C., Blasius, J., Durrand, C. (eds.) Recent Developments and Applications in Social Research Metho-dology – Proceedings of the RC 33 Sixth International Conference on Social Science Methodology. Verlag für Sozialwissenschaften, Wiesbaden (2005) (CD-Rom)
2. Alisch, L.-M.: Multimodale Situationsanalyse (MSA). Entwicklungsstand und Probleme. In: Beck, K., Zlatkin-Troitschanskaia, O. (eds.) Lehrerprofessionalität. Was wir wissen und was wir wissen müssen, Lehrerbildung auf dem Prüfstand 2010, Sonderheft (Multi-modal Scene Analysis: Prospects and Problems), pp. S71–S85 (2010)

3. Anderson, P.A., Anderson, J.F.: Measurements of Perceived Non-verbal Immediacy. In: Manusov, V. (ed.) The Sourcebook of Non-verbal Measures, pp. 113–126. Erlbaum, Mahwah (2005)
4. Anderson, P.A.: Nonverbal Communication: Forms and Functions. McGraw Hill, Boston (1999)
5. Babad, E.: Nonverbal Behaviour in Education. In: Harrigan, J.A., Rosenthal, R., Scherer, K.R. (eds.) The New Handbook of Methods in Nonverbal Behaviour Research, pp. 283–311. Oxford University Press, Oxford (2005)
6. Balduzzi, D., Tononi, G.: Qualia: The Geometry of Integrated Information. PLOS Computational Biology 5(8), e1000462, 1–24 (2009)
7. Beck, K.: Die empirischen Grundlagen der Unterrichtsforschung. Hogrefe, Göttingen (Empirical Foundations of Instructional Research) (1987)
8. Bejan, A.: Shape and Structure from Engineering to Nature. Cambridge University Press, Cambridge (2000)
9. Bejan, A.: The Constructal Law in Nature and Society. In: Bejan, A., Merkx, G.W. (eds.) Constructal Theory of Social Dynamics, pp. 1–33. Springer, New York (2007)
10. Brugman, C.: Story of Over. M.A. Thesis. Berkeley, University of California (1981)
11. Clark, A., Lappin, S.: Linguistic Nativism and the Poverty of Stimulus. Wiley-Blackwell, Chichester (2011)
12. Dawson, M.R.W.: Minds and Machines. Connectionism and Psychological Modeling. Blackwell, Malden (2004)
13. Domotor, Z.: Measurement from Empiricist and Realist Points of View. In: Savage, C.W., Ehrlich, P. (eds.) Philosophical and Foundational Issues in Measurement Theory, pp. 195–221. Erlbaum, Hillsdale (1992)
14. Domowitz, J., El-Gamal, M.A.: A Consistent Nonparametric Test of Ergodicity for Time Series with Applications. Journal of Econometrics 102, 365–398 (2001)
15. Dretske, F.: Knowledge and the Flow of Information. MIT Press, Cambridge (1981)
16. Feyerabend, P.: Consolations for the Specialist. In: Lakatos, I., Musgrave, A. (eds.) Criticism and the Growth of Knowledge, pp. 197–230. Cambridge University Press, Cambridge (1970)
17. Field, H.: Theory Change and the Indeterminacy of Reference. The Journal of Philosophy 70, 462–481 (1973)
18. Fritz, P.K., Victoir, N.B.: Multidimensional Stochastic Processes as Rough Paths. Cambridge University Press, Cambridge (2010)
19. Gao, J., Cao, Y., Tung, W.-W., Hu, J.: Multiscale Analysis of Complex Time Series. Wiley, Hoboken (2007)
20. Grice, H.P.: Meaning. Philosophical Review 66, 377–388 (1957)
21. Hacking, I.: Representing and Intervening. Introductory Topics in the Philosophy of Natural Science. Cambridge University Press, Cambridge (1983)
22. Hempel, R., Westfeld, P.: Statistical Modeling of Interpersonal Distance with Range Imaging Data. In: Esposito, A., Hussain, A., Marinaro, M., Martone, R. (eds.) COST Action 2102. LNCS, vol. 5398, pp. 137–144. Springer, Heidelberg (2009)
23. Johannes, J.: Nonparametric Estimation in Functional Linear Model. In: Dabo-Niang, S., Ferraty, F. (eds.) Functional and Operational Statistics, pp. 215–221. Physica, Heidelberg (2008)
24. Kuhn, T.S.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago (1962)
25. Kuhn, T.S.: Afterwords. In: Horwich, P. (ed.) World Changes: Thomas Kuhn and the Nature of Science, pp. 311–341. MIT Press, Cambridge (1983)

26. Lakoff, G.: Cognitive Semantics. In: Eco, U. (ed.) Meaning and Mental Representations, pp. 119–154. Indiana University Press, Bloomington (1988)
27. Lee, J., Marsella, S.C.: Nonverbal Behavior Generator for Embodied Conversational Agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)
28. Lewis, D.: Convention. Harvard University Press, Harvard (1969)
29. Lewis, D.: Languages and Language. In: Gunderson, K. (ed.) Minnesota Studies in the Philosophy of Science VII, pp. 3–35. University of Minnesota Press, Minneapolis (1975)
30. Lewis, D.: Convention, Reissued edn. Blackwell, Oxford (2002)
31. Lindner, S.: A Lexico-Semantic Analysis of Verb-Particle Constructions with Up and Out. Ph.D. Dissertation. University of California, San Diego (1981)
32. McCroskey, J.C., Richmond, V.P., McCroskey, L.L.: Nonverbal Communication in Instructional Contexts. In: Manusov, V., Patterson, M.L. (eds.) The SAGE Handbook of Nonverbal Communication, pp. 421–436. Sage, Thousand Oaks (2006)
33. Müller, W.G.: Collecting Spatial Data, 3rd edn. Springer, Berlin (2007)
34. Papavasiliou, A., Ladroue, C.: Parameter Estimation for Rough Differential Equations, CRISM Paper No. 09-01 (2008), http://www.warwick.ac.uk/go/crism
35. Paulukonis, S.T., Myford, C.M., Heller, J.J.: Formative Evaluation of a Performance Assessment Scoring System. In: Wilson, M., Engelhard Jr., G. (eds.) Objective Measurement. Theory Into Practice, vol. 5, pp. 15–40. Ablex, Stamford (2000)
36. Perovich Jr., A.N.: Inkommensurabilität - ihre Unterarten und ihre ontologischen Konsequenzen. In: Duerr, H.P. (Hrsg.) Versuchungen. Aufsätze zur Philosophie Paul Feyerabends. 2. Band, Suhrkamp, Frankfurt a. M (Incommensurability – Its Sub-species and Its Ontological Consequences), pp. S76–S94 (1981)
37. Ramsay, J.: Curve Registration. In: Ferraty, F., Romain, Y. (eds.) The Oxford Handbook of Functional Data Analysis, pp. 235–258. Oxford University Press, Oxford (2011)
38. Rohwer, C.: Nonlinear Smoothing and Multiresolution Analysis. Birkhäuser, Basel (2005)
39. Scheibe, E.: Die Reduktion physikalischer Theorien. Teil II: Inkommensurabilität und Grenzfallreduktion. Springer, Berlin (1999) (The Reduction of Physical Theories. Part II: Incommensurability and Limit Case Reduction)
40. Skyrms, B.: Signals. Evolution, Learning & Information. Oxford University Press, Oxford (2010)
41. Steels, L.: Evolving Grounded Communication for Robots. Trends in Cognitive Sciences 7(7), 308–312 (2003)
42. Steels, L., Balpaeme, T.: Coordinating Perceptually Grounded Categories Through Language: A Case Study for Colour. Behavioural and Brain Sciences 28, 469–529 (2005)
43. Stevens, S.S.: Measurement, Statistics, and the Schemapiric View. Science 161, 849–856 (1968)
44. Tomasello, M.: Origins of Human Communication. MIT Press, Cambridge (2010)
45. Tononi, G.: Consciousness as Integrated Information: A Provisional Manifesto. Biological Bulletin 215, 216–242 (2008)
46. von Sachs, R., Neumann, M.H.: A Wavelet-based Test for Stationarity. J. Time Ser. Anal. 21, 597–613 (2008)
47. Wang, X.: Incommensurability and Cross-Language Communication. Ashgate, Burlington (2007)
48. Westfeld, P., Hempel, R.: Range Image Sequence Analysis by 2.5-D Least Squares Tracking With Variance Matrix Estimation. Intern. Arch. of Photogrammetry, Remote Sensing and Spatial Information Sciences B5 37, 457–462 (2008)

A Conversation Analytical Study on Multimodal Turn-Giving Cues

End-of-Turn Prediction

Ágnes Abuczki

University of Debrecen, Department of General and Applied Linguistics, Debrecen, Hungary
abuczki.agnes@gmail.com

Abstract. The present paper focuses on the systematic study of the sequential organization of verbal as well as nonverbal behavior in spontaneous interaction. The study concerns one of the most universal structural features of conversation, the phenomenon of speaker change, as occurring in forty-four dialogues of the multimodal HuComTech corpus of Hungarian spontaneous speech. The purpose of the paper is twofold: (1) to capture salient communication patterns and organized structures across the conversations, and (2) to make explicit the simultaneously occurring markers and cues of the turn-giving intention of the current speaker based on information coming from different modalities, involving: (a) verbal-acoustic (duration of continuous speech), (b) nonverbal-acoustic (duration of pauses), and (c) nonverbal-visual (gaze direction, hand gestures, posture) information. Performing several SQL queries on the HuComTech database of manually annotated spontaneous dialogues will help us determine the multimodal features of turn-ends in Hungarian. The final goal is to contribute to the development of dialogue management systems with a decision tree distinguishing two basic discourse segments, turn-keep and turn-give.

Keywords: multimodality, human–computer interaction, conversation analysis, turn management, end-of-turn prediction.

1 Introduction

The basic theoretical assumption behind the research is that owing to intrapersonal adaptation, the verbal and nonverbal behavior of a speaker is orchestrated together in a somewhat predictable manner. The present study can be placed into the research program of conversation analysis outlined by Sacks, Schlegoff and Jefferson who first shed light on the fact that everyday conversation is a structurally organized and temporally-coordinated joint activity. Instead of focusing on abstract grammatical units, such as phrases or sentences, they defined the basic units of talk from the perspective of the actual speakers [1]. The units of the interaction analysis are the ones used and allocated by the speakers, that is, their floor allocation practices: their turns.

The systematic description of the basic structure of multimodal interpersonal communication, involving turn management, is indispensable for the modeling of human–computer interaction since the way people take turns in a synchronized

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 335–342, 2012.

manner seems to be the most salient feature of interpersonal communication. As Sacks points it out: "They [turn-takings] hold across types of conversations – arguments, business talks, whatever else. They hold across the parts of a conversation – beginnings, middles, ends. They hold across topics." [2]. Therefore, one of the fundamental requirements of a dialogue system is the ability to predict the end-of-turn of the speaker with accurate timing, so that the computer agent can start his adjacent turn. A further question may arise: why do we need a multimodal approach in communication modeling? First of all, the simplest answer lies in the fact that the more modalities we have available, the safer predictions we can make. Secondly, because "the mechanisms by which people take turns in discourse are not just verbally regulated; these processes are spoken and nonverbal, as well as open and subconscious" [3]. From all the above mentioned means of conversation regulation (spoken and nonverbal processes), the goal of our research is to uncover the machine detectable acoustic and visual features of turn-ends in Hungarian dyadic interviews (without the comprehension of the lexical meaning of the utterances).

2 Research Material, Methods and Goals

The present study approaches the phenomenon of speaker change from both a theoretical and an empirical perspective. Concerning Hungarian language, due to the lack of manually annotated multimodal corpora before the compilation of the HuComTech database, the systematic, quantitative and multimodal study of conversation was not possible earlier. The field of computational pragmatics and dialogue modeling is rather new in Hungary, as opposed to Western Europe and North America where the traditions and computational practices of conversation analysis and dialogue management systems are much richer. Following the corpus-based approach in communication modeling, we examined speaker changes in a representative multimodal corpus of Hungarian spontaneous speech, the HuComTech database which involves formal job interviews (10–12 minutes each) and informal dyadic conversations (14–16 minutes each) of 113 speakers, annotated at several levels: (a) audio, (b) syntactic, (c) video, and (d) pragmatic levels, with several further sub-levels. For a detailed description of the annotation system of the HuComTech corpus, see Papay's publication who overviews the underlying theoretical assumptions and the annotation guidelines of the HuComTech scheme [4].

From the video annotation levels – including the possible physical markers of speaker changes and many other phenomena –, we would like to highlight the levels of gaze directions, posture types, and hand gestures since these nonverbal-visual markers play a crucial role in turn regulation. From the five different audio levels annotated in the Praat speech analyzer software [5], we would just like to focus on the discourse level in which the following discourse units are segmented and labeled: turn-take (TT), turn-keep (TK), turn-give (TG), backchannel (BC), and silence (SL, at least 250 ms long). Since one of the long-term goals of the HuComTech project is the development of a job interview dialogue management system demo, which must include speaker change prediction, the most important discourse segment types to distinguish are turn-give (enabling speaker change) and turn-keep (so that the computer agent can start either backchanneling or performing its next turn).

In order to complement the corpus-based approach with rule-based methods, we have tried to define a few rules and procedures that can determine and predict the occurrence of a possible speaker change (also called transition relevance places). As a starting point in our research, we used the rule-based, unimodal end-of-turn predicting model of Troung and his colleagues (see Fig. 1) who collected the major acoustic features around the time of speaker change: (a) pause length preceding speaker change, (b) the duration of speech before pause, (c) pitch contour during speech, (d) pitch change around the closing phase of speaking, and (e) the length of the prominent pitch change [6].

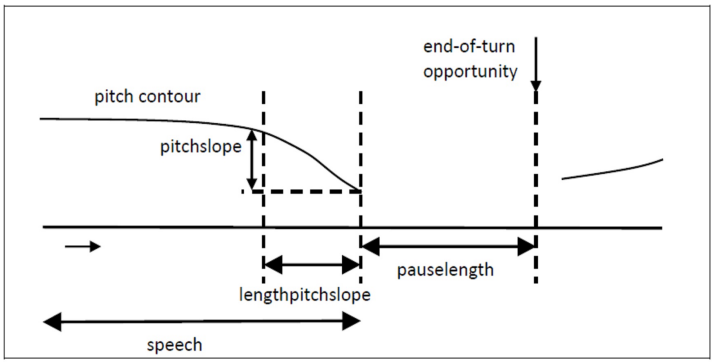


Fig. 1. The audio-based end-of-turn opportunity predicting model of Troung and his colleagues [6]

We run SQL queries to find out the number of occurrences and durations of the above overviewed acoustic features in the HuComTech database in order to determine the typical duration of these phenomena and reveal the sequential organization of interactions. Annotation files of the HuComTech corpus are stored in an SQL-database which enables us to perform various kinds of queries. Among others, horizontal label statistics – revealing the sequential organization of labels – measure the min, max, and average duration of silences before speaker change as well as the min, max, and average duration of continuous speech preceding these silences. Our database queries aimed at supplementing Troung and his colleagues' audio-based model with visual features of transition-relevance places. In order to build the outlines of a multimodal end-of-turn opportunity predicting model, we found it necessary to involve two nonverbal-visual cues in our hypothesis that are of fundamental importance in regulating the allocation of turns: (a) gaze direction, and (b) the presence or cessation of manual gesturing. Therefore, vertical label statistics were also performed in order to reveal the co-occurrences of the audio-based turn-give label with labels from the visual domain such as certain gaze direction, hand shape, and posture types.

3 Queries and Results

The queries described were performed on the cross-checked annotations of the formal and informal recordings of 22 speakers (10 male, 12 female), involving 44 .qnt files

(22 video annotations of the formal, and 22 video annotations of the informal s cenario) and 44 Praat textgrids (22 audio annotations of the formal, and 22 audio annotations of the informal scenario) temporally synchronized. As it is outlined above, we first examined the role of silence (SL) in turn regulation both qualitatively and quantitatively. As it was expected, we found that pauses within the turn of a speaker (in our terminology, during turn-keep) are generally shorter (0,31 s on average, ranging from 0,25 s to 1,8 s) than pauses right before the starting point of turn-take (pauses at transition relevant places at the end of a turn). (It should be mentioned again that only pauses longer than 0,25 s are labeled as SL in our annotation system.) As it can be read in Fig. 2, the average duration of silence preceding turn-take is approximately 500 ms in the HuComTech corpus (with 0,27 s as min, and 3,77 s as max duration). Concerning the standard deviation, and the average duration of pauses before speaker change, no significant difference was found between males and females.

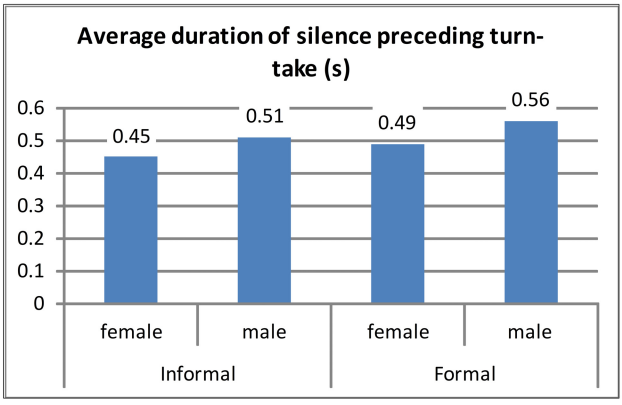


Fig. 2. Average duration of silence preceding speaker change

Then we measured the min, max, and average duration of continuous speech before speaker change (speaker change is labeled turn-take in the HuComTech annotation scheme) or continuous speech (labeled turn-keep in the HuComTech database) preceding the pauses before speaker change. According to our annotation guidelines, we consider a speech segment continuous speech if it is not interrupted with a pause longer than 250 ms. As shown in Fig. 3, the average duration of continuous speech during turn-keep among the formal and informal interviews of male and female participants varies between 2,7 s and 3,5 s; however, there are huge deviations among the individual durations, ranging from 0,3 s to 15,8 s.

As for the visual features of turn-give segments, our hypothesis is that gaze direction and the cessation of manual gesturing are two of the most typical nonverbal features in end-of-turn prediction. In 76% of all the turn-give segments (in other words, end-of-turn opportunity) examined (in 44 interviews), no hand gesture label was attached to the turn-give discourse segment. In contrast, turn-keep segments without a single hand gesture label make up only 41% of all turn-keeps. We may conclude then that speakers are much more likely to gesture within their turn than at the end of their turn.

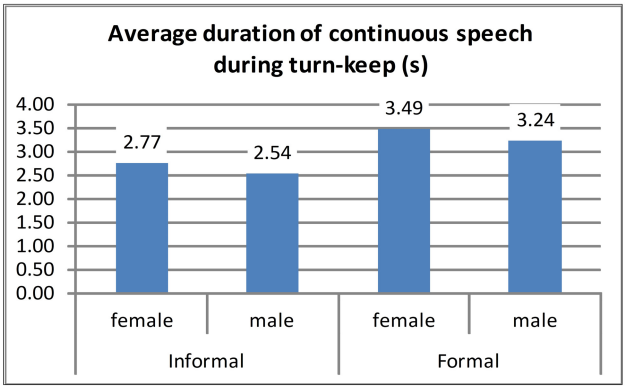


Fig. 3. Average duration of continuous speech during turn-keep

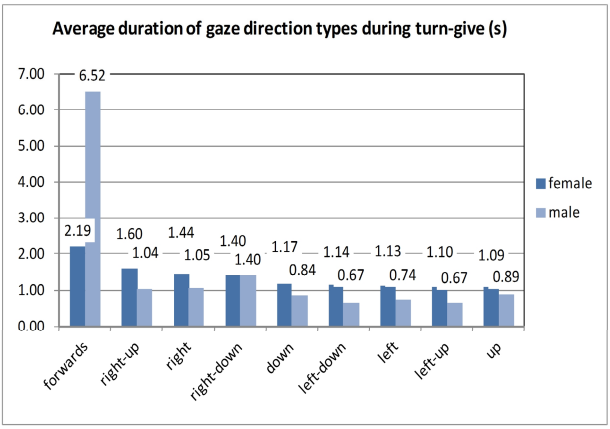


Fig. 4. Average duration of gaze direction types before speaker change

As it can be seen in Fig. 4 and Fig. 5, looking forwards is the most common nonverbal speaker behavior both in terms of its duration and frequency during the turn-give discourse segment, that is, at the end of a turn. Evidenty, maintaining long eye contact (between 2 s and 7s) - especially making eye contact without speaking - signals a transition relevance place where speaker change is most probable. On the contrary, the direction of gaze is relatively quickly changing during turn-keep since gaze labels (including all its sub-types) within a turn are 2,05 s long on average. Surprisingly, looking away from the interlocutor (all gaze label types, excluding the label 'forward') has often coincided with the beginnings of turns (in 63% of all cases of turn-takes examined). A possible explanation of speakers' tendency to often look sideways while speaking is that they try to avoid information overload.

In general, looking at the conversational partner or looking away from the partner can provide indirect cues of the speaker's willingness to continue interaction, and gazing at particular elements in the vision field can tell where the speaker's focus of attention is [7], [8].

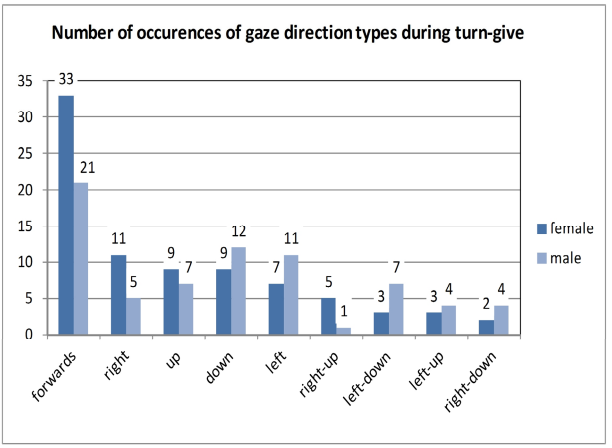


Fig. 5. Number of occurrences of gaze direction types before speaker change

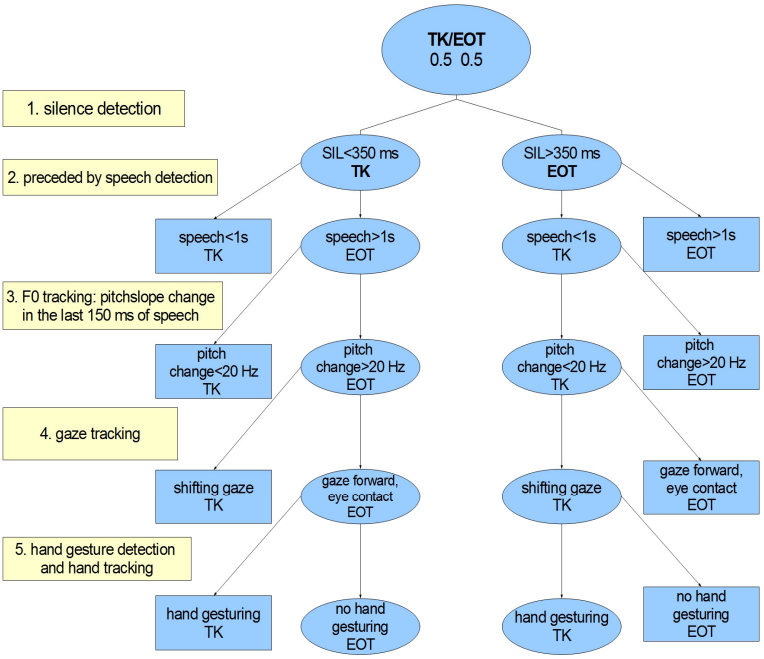


Fig. 6. A decision tree with multimodal features distinguishing the turn-keep (TK) of a speaker and end-of-turn (EOT) opportunity

4 Conclusions

In sum, it is argued in this study that multimodal approach is indispensable in communication modeling in order to disambiguate the actual meaning of polysemous communicative signals such as pauses. On the one hand, a speaker may pause within his turn; on the other hand, a pause may be a marker of a transition relevance place where speaker change may or should occur. Therefore, it has been demonstrated that communicative components from other (nonverbal) modalities are also needed, such as body movements and eye contact, in order to make explicit the actual communicative function of a pause. As a result, instead of considering communication as a predominantly verbal activity, and examining verbal utterances (speech acts) in spontaneous interaction, the entire multimodal human behavior with its composite utterances should be observed [9], [10].

Based on qualitative observations, we have found that the most reliable nonverbal cue for turn-give is the gaze behaviour of the present/previous speaker which can best be described as a long glance. Typically, the current floor holder continuously (longer than earlier) looks at the other conversation partner, seeking a reply, a future action or any reaction. Gaze shifts toward the listener frequently coincide with a shift in conversational turn – they can be seen as a signal that the floor is available. As far as the phenomenon of gesturing is concerned, no gesturing can usually be observed during this discourse segment; the cessation of manual gesturing is especially typical of turn-give. If the current floor holder spends a considerable percentage of time gesturing at the end of an utterance, he/she may be more likely to continue controlling the floor after the end of the utterance. Posture shifts also tend to occur at discourse boundaries and speaker changes. Regarding acoustic features, the most typical intonation pattern of the turn-give segment in Hungarian is falling question intonation with decreasing pitch or loudness, followed by silence [7], [8].

Based on the cross-modal queries, the following set of features was found to describe the canonical case of turn-ends in interviews (shown in Fig. 6):

- a pause of at least 300 ms (500 ms on average)
- preceded by continuous speech of at least 1 s (3 s on average)
- where the duration of the last pitch change is between 0,05–0,2 s
- which involves pitch change (usually fall) of at least 20–30 Hz
- with the head and gaze of the speaker turning towards the listener (making eye contact with the conversation partner)
- without performing hand gestures.

Concerning the decision tree shown in Fig. 6, no actual algorithm has been written up to the present date using this decision tree; however, the reliability of the decision tree has been manually checked: analyzing, measuring and contrasting the acoustic and visual features 30 randomly selected turn-keep segments and 30 turn-give/end-of-turn segments from the perspectives of the above outlined factors (the 5 steps/decisions themselves in the decision tree). Depending on the position of the end node, the rate of reliability varies between 65–90%. The further stage the end node is in (such as in step 4 or step 5), the more reliable the result is. Of course, its reliability will be

improved in the future by extending it with further multimodal features and by individually fine-tuning the figures (duration values) according to the average speech tempo of the actual speaker.

As a result, with the help of applying available pause detection, hand tracking, and gaze tracking software, the present study with its decision tree may contribute to the development of dialogue management systems with the automatic prediction of the turn-ending and turn-giving intention of the current speaker.

Acknowledgements. The database construction is a part of the Theoretical fundamentals of human-computer interaction technologies project (TAMOP-4.2.2- 08/1/2008-0009). The publication is supported by the TÁMOP 4.2.1./B-09/1/KONV-2010-0007 project. The project is co-financed by the European Union and the European Social Fund.

References

1. Sacks, H., Schlegoff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking in conversation. Language 50(4), 696–735 (1974)
2. Sacks, H.: Lectures on Conversation. Blackwell, Oxford (1992)
3. Wiemann, J.M., Knapp, M.L.: Turn-taking in Conversation. Journal of Communication, 75–92 (1975)
4. Papay, K.: Designing a Hungarian Multimodal Database – Speech Recording and Annotation. In: Esposito, A., Esposito, A.M., Martone, R., Müller, V.C., Scarpetta, G. (eds.) COST 2102 Int. Training School 2010. LNCS, vol. 6456, pp. 403–411. Springer, Heidelberg (2011)
5. Boersma, P., Weenink, D.: Praat: Doing Phonetics by Computer 5.1.43 (2010), http://www.praat.org
6. Troung, K.P., Poppe, R., Heylen, D.: A rule-based backchannel prediction model using pitch and pause information. In: Proceedings of Interspeech, Makuhari, Japan, pp. 3058–3061 (2010)
7. Abuczki, Á.: Multimodal Annotation and Analysis of Turn Management Strategies. In: Balogné Bérces, K., Földváry, K., Mészárosné Kóris, R. (eds.) HUSSE10-Linx. Proceedings of the HUSSE 2010 Conference Linguistics, pp. 107–115. Hungarian Society for the Study of English, Debrecen (2011), http://mek.oszk.hu/10100/10172/
8. Abuczki, Á.: A multimodal analysis of the sequential organization of verbal and nonverbal interaction. Argumentum 7, 261–279 (2011), http://argumentum.unideb.hu/2011-anyagok/works/AbuczkiA.pdf
9. Enfield, N.J.: The Anatomy of Meaning. Speech, gesture, and composite utterances. Cambridge University Press, Cambridge (2009)
10. Kendon, A.: Gesture. Visible Action as Utterance. Cambridge University Press, Cambridge (2004)

Conversational Involvement and Synchronous Nonverbal Behaviour

Uwe Altmann[1], Catharine Oertel[2,3], and Nick Campbell[2]

[1] Institute of Psychology, Friedrich-Schiller-University Jena, Germany
[2] Speech Communication Lab, Trinity College, Dublin
[3] Department of Speech, Music and Hearing, KTH, Sweden

Abstract. Measuring the quality of an interaction by means of low-level cues has been the topic of many studies in the last couple of years. In this study we propose a novel method for conversation-quality-assessment. We first test whether manual ratings of conversational involvement and automatic estimation of synchronisation of facial activity are correlated. We hypothesise that the higher the synchrony the higher the involvement. We compare two different synchronisation measures. The first measure is defined as the similarity of facial activity at a given point in time. The second is based on dependence analyses between the facial activity time series of two interlocutors. We found that dependence measure correlates more with conversational involvement than similarity measure.

Keywords: nonverbal behaviour, facial activity, synchronisation, involvement, windowed cross-lagged regression.

1 Introduction

In interpersonal interaction we can observe phenomena such as synchronous movements, imitation of facial expressions and gestures, posture mirroring, and convergence of prosodic cues. All can be summarized under the term synchronisation phenomena [1], [2], [9], [14].

The interdisciplinary synchronisation research focuses mainly on interactions between two individuals (called: dyadic interactions). Developmental psychology is, for instance, concerned with investigating the relationship between the synchronisation in mother-baby interactions and the subsequent development of children [9], [28]. Clinical research focuses on the relationship between synchronisation in the patient-psychotherapist interactions and the therapeutic success [20], [26], [27], [29]. Education research examines the relationship between synchronisation in teacher-student interactions and learning outcome [15], [18]. Friendship research assumes that the synchronisation of nonverbal behaviour reflects that friends are "on the same wavelength" [2], [10], [13].

Overall, research suggests a correlation between the frequency of synchronisation phenomena and the quality of interaction in the sense of the predominance of well-being, unity and solidarity [2], [4], [12]. Motivated by these findings, computer scientists are trying to make avatars synchronize their nonverbal behaviour

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 343–352, 2012.

to those of their human interlocutors in order to increase the naturalness of the interaction [6], [16], [17], [19].

1.1 Synchronisation of Nonverbal Behaviour and Its Measurement

Measuring synchronisation of nonverbal behaviour can be carried out in various ways. One possibility is a perceptual rating of an interaction sequence based on personal impressions such as similarity, mutual reference, well-ordering, or smooth-flow of nonverbal behaviour of the interlocutors (see e. g. [3]). In this case, synchronisation is defined as an observer's impression and refers (mostly) to more than one behaviour modality.

Other definitions are based on investigations of time series analyses of single behaviour modalities. Analysis is, for instance, carried out by computing the difference of a personal characteristic for each point in time. Thus, synchronisation is defined as a similarity measure for a given point in time. Such an approach is mainly found in the context of voice analyses (called "convergence of prosodic cues", see e. g. [6], [7]) and less frequently for motion analyses (see e. g. [23]).

A third approach defines synchronisation as a (temporary) linear dependence between the time series of interaction partners. In contrast to other approaches, the relationship between the time series sections of one interlocutor and delayed time series sections of the second interlocutor are investigated. Such an approach to synchronisation can be found mainly in the area of motion analyses (see e. g. [1], [11], [26]).

Considering the diversity in approaches to quantifying synchronisation the question arises whether one approach should be preferred over the other. In order to answer this question we examine the correlation between interaction quality and different synchronisation measures. We hypothesise that the "better" approach should have a stronger correlation with our measurement of interaction quality.

1.2 Conversational Involvement

How can we measure the quality of interaction? One possibility is a rating of conversational involvement. Involvement correlates with the degree of interest as well as attentiveness in a conversation [32, p. 1330]. A high degree of involvement is characterised by closer distance, mutual gaze, greater facial expressiveness, forward lean, a higher degree of gesturing, and vocal expressiveness [5, p. 463], [25, p. 53]. A low degree of involvement corresponds, for example, to avoiding mutual gaze, increasing the interpersonal distance, or turning the body away from the interlocutor (ibid).

In research concerned with dyadic interactions, involvement is usually conceptualized as a characteristic of one person (see e. g. [5], [13], [25], [32]). In dyadic interactions it is quite clear that one's involvement behaviour is addressed to the interaction partner. In case of a multi-party interaction, this methodology is problematic as, for instance, subgroups could emerge temporarily. Moreover, one person may choose not to contribute for a given time interval but rather focus on a different activity. In such cases it is not clear which interaction partner

is addressed (on one specific person, on two persons, or on all persons?). For this reason we define "conversational involvement" acording to Oertel [7], [21], [23], [24] as a group characteristic which can change over time. Conversational involvement refers to the behaviour of all participants present within a specific time interval.

It should be noted that the concept of involvement and synchronisation partially overlap in their semantic definitions. Coker and Burgoon [5] and Guerrero [13], for example, conceptualize involvement as a five dimensional variable. The dimension "immediacy" refers to typical behaviours such as mentioned above. The dimension "interaction management" can, in a broader sense, be interpreted as synchronisation as it is defined as "the degree to which participants in conversation engage in smooth-flowing conversation" [5, p. 463]. Therefore, a conversational involvement rating comprises both the degree of activity as well as the synchronisation in behaviour between participants. From a conceptional point of view both conversational involvement and synchronisation should correlate.

1.3 Research Question

In a first step, we aim at investigating and comparing various measures of synchronisation by correlating them to interaction quality. From a conceptional point of view, the "better" synchronisation measure should be correlated more with interaction quality. As a measure of interaction quality we use ratings of conversational involvement. Our sample consists of two group interactions which are, for the most part, dyadic in nature. Therefore, our analysis focuses on time series analyses of the facial activity of those two interlocutors. In a second step, based on the prefered synchronisation measure we explore the time course of the synchronisation frequency.

2 Methods

2.1 Sample, Ratings, and Motion Capture

We study two-thirty-minute-interactions which are part of the D64 corpus [22]. The same five participants are present in both interactions. However, large parts of the interaction, particularly Session 2, are mainly dyadic [23].

Based on the video recordings of the interactions, the conversational involvement is rated for every five second interval. The rating is developed by [21]. The 10 involvement levels are described in Table 1. It should be noted that what we consider "conversational involvement" in this study. Therefore, we only measure involvement when people are participating in dialogue. We do not account for the involvement of an individual person but $n > 2$ needs to be fulfilled. Our annotators account for all participants in a conversation. If, for example, in a multi–party conversation with five people two are actively participating and three look and act bored, annotators cannot only take the actively participating people into account but also need to take the other three into account when labelling the data for involvement.

Table 1. Description of involvement levels (see [21], [23], [24])

Level	Description
1	virtually no interaction, persons are not taking notice of each other and are engaged in completely different pursuits
2	less extreme variant of involvement level 1
3	should annotated when subgroups emerge
4	only one conversation is taking place
5	persons show mild interest in the conversation
6	persons encourage the turnholder to carry on
7	persons show increased interest and actively contribute to the conversation
8	persons contribute even more actively to the conversation
9	persons show absolute, undivided interest in the conversation and each other and vehemently emphasise the points they want to make; persons signal that they either strongly agree or disagree with the turn-holder
10	level 10 is an extreme variant of involvement level 9

Furthermore, with the algorithms of [30] and [31] we measure the intensity of facial activity. This is carried out by using the video data coordinates of the faces at each frame composed by the exact spot of the top left corner and the bottom right corner of the face.Normalisation is carried out as these coordinates are highly dependent on the distance of the person to the camera.

We obtain two time series ($X_{1\,t}$ and $X_{2\,t}$) with each 25 values per second per interaction. The time series were subjected to a Box-Cox-transformation with $\lambda_{\text{BoxCox}} = 0$ so that we get a more variance stable time series. Futhermore, we smooth the time series (roughness penalty of smoothing splines $\lambda_{\text{penalty}} = 0.995$).

2.2 Measurement of Synchronisation

Using the time series we compute the similarity of facial activity. This is our first synchronisation measure. It is the absolute deviation of both personal time series at a given time point:

$$X_t^{\text{diff}} = |X_{1\,t} - X_{2\,t}| \quad . \tag{1}$$

For the identification of synchronous facial movements in terms of a dependence measure we use a windowed cross-lagged regression (WCLR, for details see [1], [2]). Firstly, temporal relationships between both time series of facial activity explored with a WCLR (window width $b = 100$ time points $= 4$ seconds). In doing so, we only consider the time lags $\tau \in \{-60, \ldots, 0, \ldots, +60\}$ between both windows. That means that the maximum time lag between one's behaviour and the predicted behaviour of the other one is 60 time points (3 seconds). Secondly, the WCLR output is analysed with a peak picking algorithm. It identifies "sync intervals" in which one times series depends significantly, and with a stable time lag, on the other one. We obtain a new times series (S_t) which indicates with one, synchrony and with zero, no synchrony at t.

2.3 Statistical Analysis

At first, we study the relationship between conversational involvement and synchronisation of facial activity. As one involvement rating refers to a 5 seconds interval, we compute the relative frequency of synchrony (F_t) for each 5 seconds interval as we do with the facial activity and the difference between both facial activities. The intervals do not overlap. The regression model for the dependence of involvement (I_t) on interaction ($I_{\text{1st interaction}}$), facial activity ($X_t^{\text{mean}}$), difference between both facial activities (X_t^{diff}), point of time (T in minutes), and frequency of synchrony (F_t) is

$$I_t = \beta_0 + \beta_1 I_{\text{1st interaction}} + \beta_2 X_t^{\text{mean}} + \beta_3 X_t^{\text{diff}} + \beta_4 T + \beta_5 F_t + \varepsilon_t \quad . \tag{2}$$

To examine predictors for the appearance of synchronisation we use the following binary logistic regression model.

$$\frac{\Pr(S_t = 1)}{1 - \Pr(S_t = 1)} = \beta_0 + \beta_1 I_{\text{1st interaction}} + \beta_2 X_t^{\text{mean}} + \beta_3 X_t^{\text{diff}} + \beta_4 T \tag{3}$$

$\Pr(S_t = 1)$ is the probability that the facial activity of both interlocutors is synchronized at t and $\frac{\Pr(S_t=1)}{1-\Pr(S_t=1)}$ is the odds ratio for synchronized vs. not synchronized at t. Because we study two interactions and the second interaction is more dyadic than the other one, we assume that the frequency of synchronisation is smaller in the 1st interaction than in the 2nd one. Therefore, we use the indicator variable $I_{\text{1st interaction}}$ as a predictor. Furthermore, we assume that the probability of synchronisation increases with the mean facial activity ($X_t^{\text{mean}} = \frac{1}{2}(X_{1\,t} + X_{2\,t})$) and decreases with the difference between both facial activities (X_t^{diff}). Finally, we assume that the probability of synchronisation increases with time (predictor: T in minutes).

Please note that in model (3) the temporal difference between t and $t+1$ is $\frac{1}{25}$ seconds. Due to the aggregation in model (2) the temporal difference between t and $t+1$ is five seconds. The time variable T is the same as t, but in both models T is given in minutes. This makes the interpretation of regression coefficient β_4 easier. In model (2), it is the increase or decrease of conversational involvement per minute.

3 Results

3.1 Evaluation of Synchronisation Measures

The results regarding the prediction of the ratings of conversational involvement are listed in Table 2. We found no significant relationship between conversational involvement and similarity measure (X_t^{diff}). All other predictors are significant. In the 1st interaction the mean rating of conversational involvement is smaller than in the 2nd interaction. The involvement increases with time and with large values of mean facial activity. As assumed ratings of conversational involvement

Table 2. Statistics of the regression with involvement as dependent variable (estimate ot regression coefficient $\widehat{\beta}$, its standard error SE, statistic T, p-value, and effect size η^2)

parameter	$\widehat{\beta}$	SE	T	p	η^2
intercept	6.285	0.108	58.437	0.000	–
$I_{\text{1st Interaction}}$	−1.053	0.075	−13.973	0.000	0.229
X_t^{mean}	0.411	0.176	2.336	0.020	0.008
X_t^{diff}	0.047	0.118	0.400	0.690	0.000
T	0.014	0.005	3.015	0.003	0.014
F_t	0.324	0.135	2.404	0.016	0.009

depend significant on the relative frequency of synchronisation (F_t). However, the effect ($\eta^2 = 0.009$) is very small[1].

3.2 Prediction of Synchronised Facial Activity

As the dependence measure indicates a higher degree of synchronisation in facial activity and a higher correlation with involvement we now investigate which features influence the occurance of synchrony. Results are reported in Table 3. The p-values suggest a relationship between synchronisation and all used independent variables. Except in the time variable the signs of all parameter estimates are in the assumed direction. The biggest effects[2] on the chance that a synchronisation of facial activity occurs are found for mean facial activity (X_t^{mean}) and difference between both facial activities (X_t^{diff}). The odds ratio for $I_{\text{1st Interaction}}$ indicates that the persons in the 2nd interaction synchronize more often than in the 1st interaction, and the odds ratio for T indicates that the tendency to synchronize slightly decreases with time.

The last result is not consistent with our hypothesis. Therefore, we plot the time course of the relative frequency of synchrony (see Figure 1). It is apparent that in both interactions the relative frequency of synchronisation changes in a nonlinear way.

One advantage of the used methods is that we get information about the dependence between persons behaviour. In Table 4 are listed the frequency of time points at which the facial behaviour of both interlocutors is synchronised and – if synchronised – at which person A predicts the behaviour of person B respectively at which person B predicts the behaviour of person A. Note, 25 time

[1] Eta-square (η^2) is an effect size measure and can be interpreted in the same way as R^2. $\eta^2 = 0.009$, for example, means that 0.9 percent of the variability in the dependent variable can be explained by the independent variable under study.

[2] Note, $\exp(\beta) < 1$ indicates a decreasing chance that the event of interest (here: synchronisation of facial activity) occurs, and $\exp(\beta) > 1$ that the chance increases. $\exp(\beta) = 1$ is tantamount to non influence of the variable under study.

Table 3. Statistics of the binary logistic regression with the odds ratio of synchronized facial activity as dependent variable (estimate of regression coefficient $\widehat{\beta}$, its standard error SE, statistic Z, p-value, and odds ratio $\exp(\widehat{\beta})$)

parameter	$\widehat{\beta}$	SE	Z	p	$\exp(\hat{\beta})$
intercept	−0.859	0.019	2541.071	0.000	0.424
$I_{\text{1st Interaction}}$	−0.101	0.015	42.967	0.000	0.904
X_t^{mean}	1.470	0.030	2429.502	0.000	4.348
X_t^{diff}	−0.445	0.019	559.076	0.000	0.641
T	−0.004	0.001	14.457	0.000	0.996

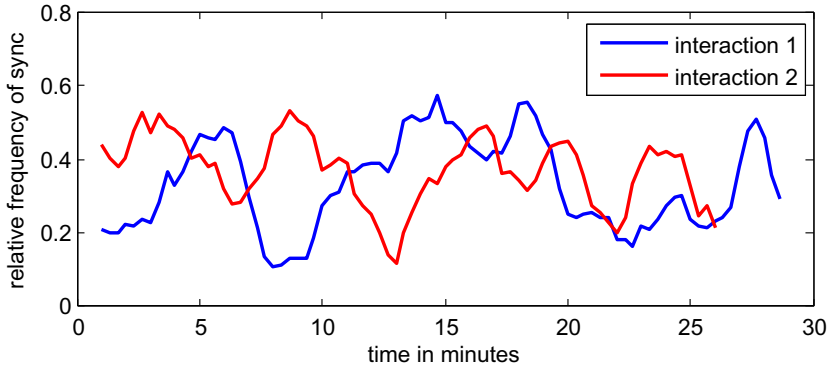


Fig. 1. Time course of the relative frequency of synchrony (moving average bases on 2 minute intervals, overlapping of sections: half a minute)

points represent one second. We found that facial activity in the 2nd interaction is more often synchronised and the frequency to lead is approximately the same for both persons.

4 Discussion

With new methods it is possible to automatically evaluate nonverbal behaviour in real time. The nonverbal behaviour could be used to estimate the current quality of interaction. In our study we examined the relationship between manual ratings of conversational involvement and two automatically extracted synchronisation measures. We chose to use conversational involvement as a measure of interaction quality as it incorporates the degree of coordination of all participants present. We used two measures to quantify synchronisation, both are based on the facial activity of two interlocutors. The first measure quantifies the similarity of facial activities of the respective interlocutors and the second indicates whether the time series section of the first person depends on a delayed times series section of the second person.

Table 4. Frequency of time points sync vs. no sync and person A vs. person B leads

	total	no sync	sync A or B leads	sync A leads	sync B leads
1st interaction	44876	33160	11716	5563	6153
2nd interaction	40426	27806	12620	6316	6304

4.1 Involvement and Synchronisation of Facial Activity

In our regression analysis, we found no significant relationship between conversational involvement and similarity measure. However, the dependence measure is significant. This suggests that in the field of motion, analyzing with the dependence measure is the best choice. However, the effect size of the dependence measure is small. A reason for this could be that our synchronisation measure is based on the facial activity alone, whereas, the ratings of conversational involvement refer to both visible behaviour and prosodic cues. These cues might have a stronger influence on the ratings than facial activity.

Our finding that conversational involvement ratings are significantly correlated with synchronous nonverbal behaviour is similar to the results of de Roten et al. [29]. They also report a relationship between involvement ratings and a micro-analytic approach based on the codings of gaze behaviour and body motions. The results confirm our theoretical assumptions that synchronisation is a sub-dimension of conversational involvement.

We found that the frequency of facial synchronisation (the dependence measure) increases in a nonlinear way. Similar results are reported by De Looze et al. [7] and Edlund et al. [8] regarding the convergence of prosodic cues and Nagaoka and Komori [20] regarding synchronous body movements. This suggests that the tendency of interlocutors to synchronize is a dynamic phenomenon.

However, the generalizability of our results is limited. We have only examined the facial activity of two persons within two group interactions. Nevertheless, the results are in harmony with the current state of research.

4.2 Outlook

In a future study we would like to extend our analysis of facial activity to all particpants within this multi-party interaction and in a further step include more modalities. Such as extension and possibly also an inclusion of further multimodal-multiparty corpora would increase the generalizability. In the long run we would like to test our findings on synchronous behaviour in a dialogue system. By getting an avatar to synchronise its behaviour to its human interlocutor we hope to increase the smooth flowing of interaction.

References

1. Altmann, U.: Investigation of Movement Synchrony Using Windowed Cross-Lagged Regression. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 335–345. Springer, Heidelberg (2011)
2. Altmann, U.: Synchronization of nonverbal behavior and it's identification with time series analysis – further development and application [Synchronisation nonverbalen Verhaltens –Weiterentwicklung und Anwendung zeitreihenanalytischer Identifikationsverfahren]. Ph.D. thesis, Friedrich-Schiller-Universität Jena, Jena, Germany (2012)
3. Bernieri, F.J., Reznick, J.S., Rosenthal, R.: Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant dyads. Journal of Personality and Social Psychology 54(2), 243–253 (1988)
4. Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology 76(6), 893–910 (1999)
5. Coker, D.A., Burgoon, J.K.: The nature of conversational involvement and nonverbal encoding patterns. Human Communication Research 13(4), 463–494 (1987)
6. Coulston, R., Oviatt, S.L., Darves, C.: Amplitude convergence in children's conversational speech with animated personas. In: Hansen, J., Pellom, B. (eds.) Proceedings of the International Conference on Spoken Language Processing, vol. 4, pp. 2689–2692. Casual Prod. Ltd. (2002)
7. De Looze, C., Oertel, C., Rauzy, S., Campbell, N.: Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In: Lee, W.S., Zee, E. (eds.) Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII), pp. 1294–1297 (2011)
8. Edlund, J., Heldner, M., Hirschberg, J.: Pause and gap length in face-to-face interaction. In: Proceedings of INTERSPEECH 2009, Brighton, United Kingdom, pp. 2779–2782 (2009)
9. Feldman, R.: Parent-infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. Journal of Child Psychology and Psychiatry 48(3/4), 329–354 (2007)
10. Field, T.M., Greenwald, P., Morrow, C., Healy, B., Foster, T., Guthertz, M., Frost, P.: Behavior state matching during interactions of preadolescent friends versus acquaintances. Developmental Psychology 28(2), 242–250 (1992)
11. Gottman, J.M., Ringland, J.T.: The analysis of dominance and bidirectionality in social development. Child Development 52(1), 393–412 (1981)
12. Guéguen, N., Jacob, C., Martin, J.: Mimicry in social interaction: Its effect an human judgment and behavior. European Journal of Social Sciences 8(2), 253–259 (2009)
13. Guerrero, L.K.: Nonverbal involvement across interaction with same-sex friends, opposite-sex friends and romantic partners: Consistency or change? Journal of Social and Personal Relationships 14(1), 31–58 (1997)
14. Harrist, A.W., Waugh, R.M.: Dyadic synchrony: Its structure and function in children's development. Developmental Review 22(4), 555–592 (2002)
15. Katsumata, G., Ogawa, H., Komori, M.: Evaluation of students' interests using analysis of speech-driven body movement entrainment. Technical Report of The Institute of Electronics, Information and Communication Engineers 109(27), 107–112 (2009), http://ci.nii.ac.jp/naid/110007333873/en/

16. Kopp, S.: Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication 52(6), 587–597 (2010)
17. Kupferberg, A., Glasauer, S., Huber, M., Rickert, M., Knoll, A., Brandt, T.: Biological movement increases acceptance of humanoid robots as human partners in motor interaction. AI & Society, Online First 26(4), 339–345 (2011)
18. La France, M., Broadbent, M.: Group rapport: Posture sharing as a nonverbal indicator. Group & Organization Management 1(3), 328–333 (1976)
19. Marin, L., Issartel, J., Chaminade, T.: Interpersonal motor coordination. From human-human to human-robot interactions. Interaction Studies 10(3), 479–504 (2009)
20. Nagaoka, C., Komori, M.: Body movement synchrony in psychotherapeutic counseling: A study using the video-based quantification method. IEICE Transactions on Information and Systems E91-D(6), 1634–1640 (2008)
21. Oertel, C.: Identification of cues for the automatic detection of hotspots, Universität Bielefeld (2010)
22. Oertel, C., Cummings, F., Campbell, N., Edlund, J., Wagner, P.: D64: A corpus of richly recorded conversational interaction. In: Kipp, M., Martin, J.C., Paggio, P., Heylen, D. (eds.) Proceedings of LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pp. 27–30 (2010)
23. Oertel, C., De Looze, C., Scherer, S., Windmann, A., Wagner, P., Campbell, N.: Towards the Automatic Detection of Involvement in Conversation. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 163–170. Springer, Heidelberg (2011)
24. Oertel, C., Scherer, S., Campbell, N.: On the use of multimodal cues for the prediction of involvement in spontaneous conversation. In: Proceedings of Interspeech 2011, Florence, Italy, pp. 1541–1544 (2011)
25. Patterson, M.L.: More Than Words: The Power of Nonverbal Communication. Aresta (2010)
26. Ramseyer, F., Tschacher, W.: Nonverbal synchrony in psychotherapy: Coordinated body-movement reects relationship quality and outcome. Journal of Consulting and Clinical Psychology 79(3), 284–295 (2011)
27. Rasting, M., Beutel, M.E.: Dyadic affective interactive patterns in the int ake interview as a predictor of outcome. Psychotherapy Research 15(3), 188–198 (2005)
28. Rogers, S.J., Hepburn, S.L., Stackhouse, T., Wehner, E.: Imitation performance in toddlers with autism and those with other developmental disorders. Journal of Child Psychology and Psychiatry 44(5), 763–781 (2003)
29. de Roten, Y., Fivaz-Depeursinge, E., Stern, D.J., Darwish, J., Corboz-Warnery, A.: Body and gaze formations and the communicational alliance in couple-therapist triads. Psychotherapy Research 10, 30–46 (2000)
30. Scherer, S., Campbell, N.: Multimodal laughter detection in atural discourses. In: Ritter, H., Sagerer, G., Steil, J. (eds.) Proceedings of the 3rd International Workshop on Human-centered Robotic Systems (HCRS 2009), pp. 111–121 (2009)
31. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
32. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations. In: Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP), pp. 1329–1332 (2004)

First Impression in Mark Evaluation: Predictive Ability of the SC-IAT

Angiola Di Conza and Augusto Gnisci

Department of Psychology, Second University of Naples,
Via Vivaldi, 43, 81100, Caserta, Italy
{angiola.diconza,augusto gnisci}@unina2.it

Abstract. According to the dual cognition theories, this paper explores the role of emotion and cognition processes of mark evaluations, by analyzing the effect of impulsive and reflective evaluations on the behaviour of approach toward that mark. The study tests the predictive contributions of the Single Category Implicit Association Test in the consumer psychology field, as a tool to detect the perceivers' first impression. Its ability to discriminate between consumers' evaluations of an unknown mark is tested according to four dimensions (Harmony, Dynamism, Pleasantness, Simplicity), whose correlation with the visual and graphical features of the mark are also tested. The SC-IAT ability to predict the following approach behaviour is tested together with the contribution of deliberative evaluations. The results indicate that the implicit evaluations affects the following behaviour, together with the explicit ones whose effects are mediated by the intentions. The findings are discussed in the frame of dual cognition models.

Keywords: SC-IAT, automatic evaluations, consumer psychology, approach behaviour, first impression.

1 Introduction

The contemporary research underlies the dual functioning of the human mind processing, based on two differently operating systems: the former, defined impulsive or automatic, is based on the activation of stored associative links, which are activated every time an object is encountered and works in a not very accurate way. The latter, defined reflective or deliberative, is activated only in the presence of time, cognitive and motivational resources, is based on the application of learned logical or mathematical rules and works in a more accurate way [1].

Among the others, the Reflect and Impulsive Model [2] maintains that the impulsive system is always activated in the presence of a stimulus; its activation can influence the reflective processing at different cognitive steps and is able to directly influence the behaviour via a shared activation of evaluative concepts stored in memory. Gawronski and Bodenhausen [3] applied this model to automatic and deliberative evaluations, asserting that the latter ones rest on the former, when the automatic evaluations are considered trustworthy. The evaluations in the impulsive

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 353–364, 2012.

system are quickly activated as soon as an object is perceived and they can influence the activation of the evaluations in the reflective system. In line with the hot cognition theory [4], the affect-based automatic evaluations are able to kick start the reflective processing, contributing to the formulation of deliberative evaluations. Meta-analytic studies on the relationship between implicit and explicit attitudes showed that implicit attitudes correlates more with affect-based explicit measures (such as 'feeling thermometers' or 'affective semantic differentials' [5]). Resting on these and similar considerations, many scholars conclude that automatic evaluations improve the understanding of decision making processes and the prediction of human behaviour, by linking affect-based and cognition-based evaluations of the object toward which the behaviour is going to be acted out [6]. Many researchers also focused their attention on the relationship between implicit and explicit attitudes and on the variables able to moderate this relation, indicating that the correlation between them can vary from very low to very high [7-8].

According to this theoretical framework, several studies have been conducted in different research fields, such as racism, stereotypes, attitudes, self-identity and the relationship between implicit and explicit evaluations, decision making processes and behaviour. Indirect measures and the analysis of the relationships among the cited constructs have been applied also to other research field, such as political psychology [9-10] and economy psychology or consumer psychology. Maison, Greenwald and Bruin [11] observed that implicit measures of product evaluations correlated with explicit measures of beliefs, feelings and behaviours. Brunel, Tietje and Greenwald [12] pointed out that the Implicit Attitude Test (IAT) is a tool reliable for identifying consumers' preferences both when they are and when they are not aware and prone to explicitly declare them. In a different study, Maison, Greenwald and Bruin [13] showed that the IAT provides information on automatic evaluations of brands (such as Coca Cola or Pepsi or two competing brands of yogurt or fast food restaurants) and predicts consumers' choices and their consequent behaviour. Isen, Labroo and Durlach [14] demonstrated that the positive affect induced by a product increases the cognitive performance and problem solving abilities, and that this effect is improved by knowing that the product is made by a popular and well-evaluated brand. These studies have generally been conducted using the most used paradigm for measuring implicit attitudes: the IAT. Notwithstanding its demonstrated psychometric characteristics, it needs two comparing categories, giving a relative index of preference for one over the two targets (e.g., giving a unique index which indicates if someone prefers Coca Cola over Pepsi or vice versa). In many fields, including the research on consumers' attitudes and behaviours, it is not always possible to identify a natural counterpart (opposing category) for every and each existing brand, above all when the to-be-evaluated brand is new, never seen before. Some instruments have been proposed as alternatives to the IAT, that utilize one target category and a couple of attribute categories (e.g., positive and negative). Among these instruments, the Single Category Implicit Association Test (SC-IAT) has been tested and it showed good psychometric properties [15].

1.1 The SC-IAT Procedure

In the last twenty years, various measures and procedures to assess automatic evaluations have been proposed. According to De Houwer's analysis[16], a measure is considered able to tap into an automatic evaluation when it provides an index of attitude even if the respondent is not aware of the fact that the attitude is being measured, do not have conscious access to the attitude or have no control over the measurement outcome. These features differentiate the measurement of implicit attitudes from the traditional questionnaire, generally employed to directly request to the respondent to evaluate an object (as the 'semantic differentials'). The measures employed to assess automatic evaluations do not ask directly or explicitly to express a judgment on an object (e.g., on behalf of a explicit question as "How do you evaluate this brand?"); on the contrary, they are based on tasks apparently not linked to the evaluative process (i.e., categorization tasks) and provide indexes of implicit attitudes derived by the computation of reaction time [16]. The SC-IAT represents a procedure for assessing this kind of automatic evaluations.

It consists of a two-phases categorization task administered by a PC: participants are asked to categorize stimuli concerning one target category (e.g., a brand) and stimuli corresponding to two opposite attribute labels (e.g., positive vs. negative) using two keys on the keyboard. In the first phase the target category is associated to one attribute category (e.g., brand + positive) placed on the left side of the screen and the stimuli presented at the screen centre and corresponding to these categories must be categorized using the computer key "A", while the category "negative" is posed on the right side of the screen and its stimuli, when presented in the screen centre, must be categorized with the key "L". In the second phase the combination of target and attribute categories is reversed, so that the target is now associated with "negative" (on the right side of the screen) and must be categorized with the "L" key; while "positive" stands alone and its stimuli must be categorized with the key "A".

The assumption is that participants spend less time to categorize with the same key the target and the attribute when they consider them as sharing the same evaluative connotation (e.g., both positive), more time when the opposite combination is proposed, that is when the same key is shared by concepts holding incongruent evaluative connotation. Assume that a participant holds a positive evaluation of a brand. Categorizing the brand with the same key used for categorizing "positive" will be an easier task (i.e., requiring less time) than using the same key for categorizing both the brand and "negative".

In a study, this paradigm was used to analyze the consumers' attitudes, to detect the implicit attitudes toward the well-known clothing brand GAP and its associations with explicit attitudes, past behaviour, behavioural intentions and future behaviour [17]. The results showed no correlation with the explicit attitudes and small-to-moderate correlation values with the other self-reported measures and with measures of behaviour, leading the authors to conclude that the SC-IAT can improve the exactness of the prediction of consumers' behaviours.

In the present study, the SC-IAT procedure is proposed as a tool to assess the automatic evaluations involved, together with the deliberative ones, in the cognitive

processes leading to a decision making. To synthesize, the automatic evaluations are generally recognized as affect-based and depending on visual and perceptive features of the stimulus, while the deliberative ones are considered cognition based and depending on a semantic and reflective processing. Both of them exert their influence in the decision making process leading to the behaviour, enlightening the joint contribution of the two systems (impulsive and reflective) in the human evaluative processing of to-be-evaluated stimuli.

1.2 Forming an Impression Implicitly and Explicitly

The attention of the studies on implicit evaluations in consumers' psychology has been devoted to the evaluations of widespread and well known brands or products [11-12, 14, 17]. This study, instead, aims at proposing the SC-IAT as a tool to detect the first impression of the perceivers, when they are exposed to a new, unknown mark. The underlining cognitive process is completely different. In the first case, the automatic evaluation is probably the stratification of all the preceding encounters with that brand and of the past experience with the product associated with the brand; in the second, the automatic evaluation is, of necessity, tied to graphical and perceptive cues of the brand. It is a 'first impression'.

An assumption of this study is that a first impression on the mark arises immediately and automatically in several evaluative dimensions. Some studies on the effects of the ads show that the simple exposure to an ad leads to the formation of an impression and of an evaluation even when the perceiver's attention is not consciously directed to it [18-20]. Seeing the image of a brand produces the building up of an initial impression, which can influence the following rational evaluations and the approach-avoidance behaviours toward that brand. Generally entrepreneurs are interested in creating a mark with distinctive characteristics that connote both the firm and the product, able to induce prospective consumers to approach them.

In sum, differently from previous ones, this study presents many novelties: it applies the SC-IAT to a new, never perceived mark, focusing on the 'first impression'; it widens the range of the automatic measures, not limiting to the traditional evaluative dimension of pleasantness (i.e., negative vs. positive impression).

1.3 Aims

Considering the relevance of both cognitive and emotional information processing in building an object evaluation and the link between evaluations, decision making and behaviour, the present contribution aims to test the effectiveness of the first impressions detected with the SC-IAT in predicting the following explicit evaluations, behavioural intention and consumers' behaviour.

In the cited studies, the SC-IAT revealed an useful tool to detect the participants' automatic evaluations toward well known brands or products; in the present study, we enlarge its field of application, by applying its use to the assessment of the first impression developed by the respondent toward a new, unknown mark (Gota Blanca, actually existing).

In addition to the commonly identified evaluative dimensions concerning pleasantness, we consider three further dimensions, identified as relevant by the mark producers and recognized relevant by the researchers, in order to test whether the SC-IAT is able to detect impressions formed by something more than the mere pleasantness of the mark and whether this measure improves the prediction of the following behaviours toward the firm represented by the mark.

We identify three main aims: (1) testing the predictive power of the explicit evaluations on the intention and on the behaviour; (2) testing the predictive power of the first impression on the explicit evaluations, on the intention and on the behaviour; (3) identifying the associative links between the visual and graphical characteristics of the mark and the implicit and explicit evaluations on the four attribute dimensions.

2 Method

The study is articulated into two steps: firstly, the firm owners were interviewed about the just born firm and about the message they aim to express through the mark (their brand and logo); starting from this information, four dimensions considered relevant have been identified and the perception of the potential consumers (the participants in the study) was tested via the SC-IAT and the semantic differential. Finally, items about the intention to approach the firm have been presented and, at the end of the data collection session, a behavioural index was registered as a measure of the tendency to approach the firm represented by the mark: this index was provided by the observation of the participants' decision of taking or not taking the catalogue of the firm products.

2.1 The Interview: The Story of the Firm and the Characteristics of the Mark

The two entrepreneurs (one female, one male) of a recently born, small-size, active and growing, enterprise were interviewed about their firm, how it was born, what they produce, and, particularly, about the message they mean to send to the prospective buyers of their products. The main information follows.

Fig. 1. The mark Gota Blanca (brand and logo)

Gota Blanca was born on 2007 in Naples, Southern Italy: it produces beachwear following the philosophy of combining art and fashion. Indeed, it is based on 100% made in Italy, high quality, handmade and hand-painted products, mainly for a female

target. The firm is presently expanding, by producing a line of beachwear for men and children also (more information about Gota Blanca are available at the website: http://www.gotablanca.it).

The logo recalls the Spanish name Gota Blanca (White Drop), indeed it is a green-and-white stylized drop moving downward (see Figure 1). Interviewed about the meanings associated with the chosen mark, the two entrepreneurs reported many terms, including pleasantness, harmony, dynamism, simplicity, nature, womanhood, etc. After a wide-range, deep reflection on the characteristics of the mark, taking into consideration also its perceptive features, the researchers focused on the first four features above-listed by the entrepreneurs. They were regarded as peculiar and subsequently used as attribute dichotomies in the SC-IAT procedures. Noteworthy, indeed, the SC-IAT can be used by proposing simple and immediate evaluative dichotomies and the chosen dimensions match these characteristics.

2.2 Sample

One-hundred-eighty-six university students took part in this study (23% males; mean age 24.17; s.d. = 5.87); 47% completed the study concerning the dimensions of Harmony and Dynamism ($N = 87$) and 53% completed the study regarding Simplicity and Pleasantness ($N = 90$).

2.3 Measures

The SC-IAT provides an index of the implicit evaluation of the target on a dichotomous dimension. It proves particularly useful when a natural opposite for a category target cannot be identified. Detecting the first impression toward a new mark does not require a contrasting category target (except when a well established brand is assumed by consumers as the comparison standard). In this study, four dichotomous and bipolar dimensions are considered, their attribute categories are the couples harmonious-disharmonious (well-balanced, proportionate, melodic, symmetrical, musical vs. unbalanced, disproportionate, striking, asymmetrical, noisy; $\alpha = .84$); dynamic-static (active, vital, alive, energetic, strong vs. passive, inactive, dead, apathetic, weak; $\alpha = .78$); simple-complex (easy, straightforward, linear, plain, understandable vs. uneasy, obscure, complex, complicated, incomprehensible; $\alpha = .75$); pleasant-unpleasant (beautiful, appealing, pleasing, positive, desirable vs. ugly, nasty, unpleasant, negative, undesirable; $\alpha = .83$). The target category is represented by the word “mark” and the stimuli are pictures of the mark (see Figure 1). Negative SC-IAT values indicate that the mark is evaluated as disharmonious, static, complex or disagreeable, while positive values indicate that the mark is evaluated as harmonic, dynamic, simple or agreeable.

Most striking visual and graphical characteristics of the mark. This variable was detected by one item asking: “Which feature of the logo strikes you?”, and having as response alternatives: “the colour”, “the shape”, “the style”, “the entirety”, “nothing”. Subsequently, starting from the participants’ answers to this question, five dummy variables where computed to be used in the following analyses.

Preliminary checklist. After completing the SC-IATs, the participants were asked to indicate if the mark refers or does not to a list of concepts, including the four relevant ones ("According to you, does the mark refer to the following concepts?" Simplicity, Harmony, Dynamism, Pleasantness), other ones suggested by the entrepreneurs as related to mark (Womanhood and Nature) and some others apparently not related to the mark, such as Sophistication, Rebellion, Artificiality, Strength, Rigidness, Immobility, Speed. Answers were coded 0 = No; 1 = Yes.[1]

The semantic differential. The reflective mark evaluation was assessed via a 11-items semantic differential (from 1= completely "*negative*" to 7= completely "*positive*"). The following items assessed the four dimensions: Harmony: harmonic, proportionate, symmetric (α = .78); Dynamism: dynamic, modern (α = .70) ; Simplicity: simple, easy, straightforward (α = .68); Pleasantness: positive, nice, appealing (α = .81). The indexes are computed in the same direction as the SC-IAT ones.

Behavioural intention. Seven items were asked concerning the intention to act several different behaviours potentially approaching the mark (to see its products; to buy from this firm; to deepen the knowledge about the firm and about its products; etc.).

Approach behaviour. A positive implicit and/or explicit evaluation is assumed to lead to an approach behaviour toward the positively evaluated object. At the end of the data collection session, the participants' decisions of taking (coded "1") or not taking (coded "0") the catalogue of the firm products was registered.

2.4 Procedure

Both the implicit and explicit measures have been administered by the Inquisit 2.0 software. First, each participant was informed about the aims of the study (knowing the participant's opinion about the mark of a new-constituted firm), and saw the mark. Then half of the participants completed two SC-IAT procedures having as attribute couples simple-complex and pleasant-unpleasant and the other half completed two SC-IAT procedures having as attribute couples static-dynamic and harmonious-disharmonious. The SC-IATs sequence was counterbalanced between subjects, as well as the order of presentation of each combination task.

Afterwards, the participants were asked to indicate if, in their opinion, the mark refers to the listed concepts (preliminary checklist) and answered the semantic differential scales. Finally, they were informed about the opportunity of taking the catalogue.

2.5 Data Analysis

The collected data were analysed descriptively and, successively, eight one-sample t-tests were conducted for verifying whether the observed means were different from

[1] A list of dimensions was proposed to the participants, who were asked to state whether, according to them, each dimension recalls or does not recall the dimension. The participants more frequently indicated that the mark recalled the following concepts Womanhood, Simplicity, Nature, Harmony, Dynamism. On the contrary, the other dimensions, including Rebellion, Strength, Rigidness, Artificiality, Speed, Immobility, resulted less frequently recalled by the mark.

the intermediate point of the scale, that is 0 for the implicit attitudes and 4 for the explicit ones, in order to understand if a polarized positive or negative evaluation on each implicit and explicit evaluation was detected.

Afterwards, structural equation models were preliminarily analysed to test the dimensionality of the explicit evaluations of the four dimensions and the impact of these evaluations on the intention and on the behaviour. These analyses were conducted on the full sample ($N = 186$).

Subsequently, models including the first impressions on couples of dimensions were conducted, because of the experimental design, for which half of the participants executed the Harmony and Dynamism SC-IAT and the other half executed the Simplicity and Pleasantness SC-IAT. As proposed by the R.I.M. [2], these models tested the impact of the first impressions on the explicit evaluations and on the behaviour; of the explicit evaluations on the intention and on the behaviour; and of the intention on the behaviour. Finally, a chi square analysis was conducted on the answers given by the participants to the items concerning the most striking element, in order to identify which ones were most frequently chosen; afterwards a correlation analysis was conducted in order to identify the association between specific visual and graphical characteristics of the mark (colour, shape, style, entirety and nothing) and the first impression and the explicit evaluations developed by the participants on each of the four dimensions.

3 Results

3.1 Descriptive Statistics

The descriptive statistics concerning the implicit and explicit evaluations on each dimension are printed in Table 1. The preliminary descriptive analyses have been conducted on the whole sample, not considering the distinction in subsamples executing different SC-IATs. All the SC-IAT mean values are positive and significantly higher than 0, except for Harmony, whose t-test results not significant. These results indicate that the participants' first impression derived from their exposure to the mark lead to the evaluation of the mark as dynamic, pleasant and simple. The reflective evaluations are above the middle point of the scale (4). All the t-tests conducted to compare the SC-IAT mean values with 0 (except Harmony) and the mean values on the semantic differential with 4 are significant ($p \leq .05$).

Table 1. Means and standard deviations for the SC-IAT values (first impression) and the semantic differential values (reflective evaluations) computed on the whole sample ($N = 186$), * $p \leq .05$

Dimension	Mean (Standard deviation)	
	First impression (intermediate value=0)	Reflective evaluation (range 1-7)
Harmony	0.05 (0.39)	5.25 (1.09) *
Dynamism	0.21 (0.36) *	5.31(0.36) *
Pleasantness	0.19 (0.43) *	5.61 (0.94) *
Simplicity	0.21 (0.37) *	4.93 (0.82) *

3.2 Dimensionality of the Explicit Evaluations and Their Predictive Impact

Two models have been compared: one (1) including a latent variable for each evaluations of the four dimensions and the other one (2) including one latent variable saturated by the four explicit evaluations. As printed in Table 2, the model (2) represents a model with a better fit, compared with the model (1). The four dimensions saturated as follows: Dynamism .53; Harmony .85; Simplicity .73; Pleasantness .81.

Table 2. Comparison between Model 1 and Model 2, to compare the dimensionality of the explicit evaluations

Model	X^2	Df	*p*	RMSEA	CFI	NFI	$\Delta\chi^2$	*p*
1	333.35	10	<.001	.42	.46	.46	—	—
2	36.28	8	<.001	.14	.95	.94	324.62	<.001

3.3 The Role of First Impressions and of Explicit Evaluations

The role of the first impressions was tested on the subsamples of participants who executed each couple of SC-IAT. For the Harmony and Dynamism dimensions (N = 87), conducting a one-tail test, we obtained that the first impression on both the dimensions do not impact the explicit evaluations, but significantly impact the behaviour (Harmony: β = .22; Dynamism: β = -.17); the explicit evaluations (one factor) influence the intention (β = .58) and the intention finally predict the behaviour (β = .49). The model fit is: $\chi^2(8) = 8.82$; p = .36; RMSEA = .03; NFI = .92; CFI = .99 and the R^2 are: for the intention .34 and for the behaviour .30.

Overlapping results were obtained for the Simplicity and Pleasantness dimensions (N = 99). No significant effects were observed for the Simplicity dimension, whereas the Pleasantness one showed to exert a significant influence on the behaviour (β = .24), but not on the explicit dimension, which, in its turn, influences the intention (β = .43), that, finally influences the behaviour (β = .63). The model fit is: $\chi^2(9) = 12.72$; p = .17; RMSEA = .06; NFI = .92; CFI = .97 and the R^2 are: for the intention .19 and for the behaviour .46.

3.4 Visual and Graphical Characteristics

The correlations between visual and graphical characteristics and the implicit and explicit evaluations were analysed in order to identify their association with the first impressions and with the explicit evaluations on each dimensions.

For the participants in both the Harmony-Dynamism and Simplicity-Pleasantness conditions, the analysis of the visual and graphical characteristics of the mark indicates that the "shape" is the most striking element ($\chi^2(4)$ = 24.90 and 27.01, for both p < .001; s_{shape} = 2.30 and 3.42, both p < .05; $s_{nothing}$ = -3.45 and -3.33, both p < .05).

Among the cited visual characteristics: the "colour" correlates positively with the implicit dimensions of Dynamism (r = .24; p = .02) and Pleasantness (r = .20; p = .03); the "style" correlates negatively with the explicit evaluation on the Dynamism

dimension (r = -.21; p = .05), while the "ensemble" positively correlates with the explicit evaluation of Pleasantness and Simplicity (r = .23; p = .02 and r = .22; p = .03, respectively). The variable "not striking", finally, correlates negatively with the Harmony implicit evaluation (r = -.21; p = .05).

4 Discussion and Conclusions

The present study aimed to test the ability of the SC-IAT to detect the influence of potential consumers' first impressions on four different evaluative dimensions, in order to improve the knowledge about the contribution given by emotion-based and cognition-based evaluations of a new mark to the consumers' approach behaviour. The literature on this issue indicates the implicit evaluations as affect-based and the explicit ones as reflection based. Moreover, both are indicated as intervening in the decision making process, potentially interacting at several points and jointly leading to a behaviour [2].

The descriptive analysis (means and standard deviations) indicates that all the dimensions are implicitly and explicitly polarised on the positive evaluative side (in the direction of dynamic, pleasant, harmonic and simple), with the only exception represented by the SC-IAT value for Harmony. These results indicate that, independently from whether the participants consider the mark as referring to each dimension, they tend to evaluate it (both impulsively and reflectively) as dynamic, pleasant and simple, indicating that the entrepreneurs obtained their aim to connote their mark with these characteristics, which are implicitly perceived by the participants to our research.

First of all, the dimensionality of the explicit evaluation was tested, leading to the result that the four considered evaluative dimensions contribute to the formation of an unique explicit evaluation, composed by the participants' evaluations on each attribute dichotomy. Furthermore, as expected, the explicit evaluation, deriving from the basic evaluation on each dimension, shows to influence the intention, which, in turn, determines the approach behaviour.

In line with previous studies, the implicit evaluations, corresponding in this study to the first impressions about a new, never seen before, stimulus (the mark) influence the behaviour, but do not exert an effect on the formation of the explicit evaluations. These findings lead to the conclusion that the first impressions on the selected dimensions (except Simplicity), as detected via the SC-IAT procedure, improve the exactness of the prediction of the following behaviour, but do not influence the reflective formulation of a judgment. It seems possible that the strength of this link increases when the implicit attitudes are automated, that is after several and repeated experiences with the stimulus, which determine the necessity to evaluate it several times, finally leading to the automation of the evaluation, also improving its trustworthiness [3]. We can conclude that the SC-IAT proved to represent a useful tool to detect first impressions developed on different evaluative dimensions, being able to detect the differences in participants' implicit evaluations, when they are exposed for the first time to a new, to-be-evaluated stimulus.

A further aim of this study was to identify the association between different perceptive characteristics (graphical and visual) of the stimulus and the implicit and explicit evaluations on the four evaluative dimensions. The analysis of the visual and graphical characteristics of the mark indicates that the overall evaluation of the mark was positive, as showed by the significant negative residuals obtained at the chi square test in correspondence to the answer alternative "nothing". This result indicated that a definitively small number of participants considered the mark not striking at all. Among the listed graphical and visual features of the mark, the participants more frequently choose the shape (a stylized drop) as the most striking perceptive element, but the correlation analyses indicates that it is the element less associated with the automatic and reflective evaluations of the mark. Indeed, the shape is negatively associated only with the explicit evaluation of Dynamism, whereas the ensemble (as suggested by the two entrepreneurs) is associated with explicit evaluation of Pleasantness and Simplicity, and the colour is associated with the implicit evaluation of Dynamism and Pleasantness. These findings lead to the conclusion that different perceptive characteristics are associated with different evaluative dimensions. The specific contribution of different perceptive features of a new stimuli in the building of automatic and reflective evaluations should be deepened in future researches.

Summarising, we can state that the present study give indications about a new interesting way to employ the information which can be collected via the SC-IAT procedure and indicated that more evaluative dimensions (over and above the Pleasantness one) should be considered as contributing to the development of the first impression toward a new object, being able to influence the following behaviour. Starting from these results, it seems worthy to understand how these concepts correlate between each other in the impulsive system and to delineate the cognitive processes leading to the formation of the link between the implicit and the explicit evaluation based on the same dimensions, as expected after the automation of the implicit ones [2-3].

Acknowledgement. Thank you to the two entrepreneurs, holding the firm Gota Blanca, Sula Sergi and Alessandro Manco, who let us use their mark and get the material needed for the study. Thanks also to Annamaria Esposito and Annunziata Longobardi for their help in data collection.

References

1. Smith, E.R., DeCoster, J.: Dual Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. Pers. Soc. Psychol. Rev. 4, 108–131 (2000)
2. Strack, F., Deutsch, R.: Reflective and Impulsive Determinants of Social Behavior. Pers. Soc. Psychol. Rev. 8, 220–247 (2004)
3. Gawronski, B., Bodenhausen, G.V.: Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change. Psychol. Bull. 132, 692–731 (2006)

4. Lodge, M., Taber, C.S.: The automaticity of affect for political leaders, groups, and issues: An experimental test of the Hot Cognition hypothesis. Polit. Psychol. 26, 455–482 (2005)
5. Hofmann, W., Gschwendner, T., Nosek, B.A., Schmitt, M.: What moderates implicit-explicit consistency? European Review of Social Psychology 16, 335–390 (2005)
6. Greenwald, A.G., Poehlman, T.A., Uhlmann, E., Banaji, M.R.: Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. J. Pers. Soc. Psychol. 97, 17–41 (2009)
7. Nosek, B.A.: Moderators of the relationship between implicit and explicit evaluation. J. Exp. Psychol. Gen. 134, 565–584 (2005)
8. Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., Schmitt, M.: A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. Pers. Soc. Psychol. B 31, 1369–1385 (2005)
9. Di Conza, A., Gnisci, A., Perugini, M., Senese, V.P.: Atteggiamento implicito ed esplicito e comportamenti di voto. Le Europee del 2004 in Italia e le politiche del 2005 in Inghilterra. Psicologia Sociale 2, 301–329 (2010)
10. Di Conza, A., Gnisci, A., Senese, V.P., Pagano, P., Schiavone, S.: La partecipazione politica modera l'effetto degli atteggiamenti impliciti sul voto? Uno studio sulle elezioni politiche del 2006 e 2008. Giornale Italiano di Psicologia 3, 627–648 (2011)
11. Maison, D., Greenwald, A.G., Bruin, R.H.: The Implicit Association Test as a Measure of Implicit Consumer Attitudes. Polish Psychological Bulletin 32, 1–9 (2001)
12. Brunel, F.F., Tietje, B.C., Greenwald, A.G.: Is the Implicit Association Test a Valid and Valuable Measure of Implicit Consumer Social Cognition. J. Consum. Psychol. 14, 385–404 (2004)
13. Maison, D., Greenwald, A.G., Bruin, R.H.: Predictive Validity of the Implicit Association Test in Studies of Brands, Consumer Attitudes, and Behavior. J. Consum. Psychol. 14, 405–415 (2004)
14. Isen, A.M., Labroo, A.A., Durlach, P.: An Influence of Product and Brand Name on Positive Affect: Implicit and Explicit Measures. Motiv. Emotion 28, 43–63 (2004)
15. Karpinski, A., Steinman, R.B.: The Single Category Implicit Association Test as a Measure of Implicit Social Cognition. J. Pers. Soc. Psychol. 91, 16–32 (2006)
16. De Houwer, J.: What are implicit measures and why are we using them. In: Wiers, R.W., Stacy, A.W. (eds.) The Handbook of Implicit Cognition and Addiction, pp. 11–28. Sage Publishers, Thousand Oaks (2006)
17. Karpinski, A., Steinman, R.B.: The Single Category Implicit Association Test (SC-IAT) as a Measure of Implicit Consumer Attitudes. European Journal of Social Science 7, 32–42 (2008)
18. Janiszewski, C.: Preconscious Processing Effects: The Independence of Attitude Formation and Conscious Thought. J. Consum. Psychol. 15, 199–209 (1988)
19. Janiszewski, C.: The Influence of Print Advertisement Organization on Affect Toward a Brand Name. J. Consum. Res. 17, 53–65 (1990)
20. Shapiro, S.: When an Ad's Influence is Beyond our Conscious Control: Perceptual and Conceptual Fluency Effects Caused by Incidental Ad Exposure. J. Consum. Res. 26, 16–36 (1999)

Motivated Learning in Computational Models of Consciousness

James Graham[1] and Daniel Jachyra[2]

[1] School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, USA
jg193404@ohio.edu
[2] University of Information Technology and Management Rzeszow, Poland
djachyra@wsiz.rzeszow.pl

Abstract. Much work has gone into designing and implementing agents capable of "cognitive" thought. In this paper, we give an overview of a motivated learning model and describe various ways in which we are in the process of implementing the model for simulation purposes. This work presents three different software platforms (Blender, iCub, and NeoAxis) through which an intelligent conscious agent can be implemented. This article presents the concept of a computational model of consciousness as a feature of a cognitive agent and discusses how it might be implemented/simulated.

Keywords: Machine consciousness, conscious agent, motivated learning, cognitive model, agent simulation.

1 Introduction

The purpose of this work is to give an overview of our work toward designing and implementing motivated conscious machines in simulated environments. In this paper we briefly examine consciousness, both in terms of the philosophical and machine implementation, discuss the model from which our work is derived, and discuss several simulated environments with witch we are working to implement and improve the model of consciousness referred to in this paper. The inspiration for this work is experiments with already developed and implemented motivated learning mechanisms. In such a learning approach, the agent receives pain signals from the environment and must find a way to reduce them. The agent, during interaction with environment, is able to create it's own value system. Going further we want to create a mechanism that can not only survive in a hostile environment but also be truly "aware" of what it actually does. Essentially, we believe the "value" of motivated learning is that the overall approach will help lead us toward conscious intelligent machines since one aspect that both our work and the development of consciousness seem to share is that "self-development" are integral to both.

Machine consciousness is also known as artificial consciousness or synthetic consciousness and is related to cognitive robotics and artificial intelligence. These

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 365–376, 2012.

fields have suggested various aspects of consciousness generally deemed necessary for a machine to be artificially conscious. Several functions in which consciousness plays a role were suggested by Bernard Baars [1] and other scientists. He suggested following functions: Adaptation and Learning, Definition and Context Setting, Editing, Recruiting and Control, Flagging and Debugging, Prioritizing and Access-Control, Analogy-forming Function, Decision-making or Executive Function, Autoprogramming and Self-maintenance Function, Definitional and Context-setting Function and Metacognitive and Self-monitoring Function. Professor Igor Aleksander suggested 12 principles for Machine Consciousness [2] among which are: Inner Neuron Partitioning, The Brain is a State Machine, Conscious and Unconscious States, The Awareness of Self, Representation of Meaning, Perceptual Learning and Memory, Will, Prediction, Learning Language, Learning Utterances, Emotion and Instinct. The main goal of Machine Consciousness is to define whether and how these and other aspects of consciousness can be synthesized in one consistent whole in the form of a computational model of a conscious agent.

There are many different positions on what exactly defines consciousness in a machine, likely as many as there are models for such machines [3, 4, 5, 6, 7]. Most possess some form of awareness, an ability to learn, and some level of prediction or anticipation (this includes estimation or the ability to plan). In our model, consciousness is viewed in relation to the ability and intelligence needed to build stable sensory representations and predict the results of various actions (performed by itself of otherwise observed in the environment). The following section describes the model for machine consciousness that we are using in more detail.

2 A Conscious Agent

The model of motivated learning and goal creation presented in [8] is designed to eventually be implemented in a more sophisticated agent. The motivation system will coexist with other interconnected components such as the memory system (consisting of short-term, long-term, episodic, and semantic memories), attention system, and a central executive system, among other systems. The recent work presented by Starzyk and Prasad [3] details such a system and how it might work [9].

As mentioned our goal is to create a form of “consciousness” capable building stable sensory representations and predicting the results of various actions. The model relies on motivation, sensory, and thought signals that compete for attention. Consciousness can be said to require the following components: a mechanism to acquire and represent knowledge, a mechanism for maintaining and switching attention, capability for parallel exploration, memory activation, and cognitive perception via sematic memories constructed form knowledge building and episodic memories.

2.1 Details of the Model

The cognitive system model [3] we’re using is shown in Figure 1. Components in this model influence each other in several ways. Additionally, the system is highly

parallel. When sensors, episodic or other sources, activate the sematic memory, it must resolve the competition between the various activations. Similarly, the attention switching block must choose the current "winner" for attention. Attention switching is a dynamic process, which results from competition between motivations, sensory inputs, internal thoughts, and other signals (such as noise or other unexpected stimuli). Typically, attention is thought of as being centered on visual representations of the physical world, but other stimuli such as thoughts and sounds can hold attention as well. At the core of the system is the Motivation block which directs the agent's focus, however to properly understand how it fits into the system, we also need to understand how the other components interact, which will be discussed in the following subsections.

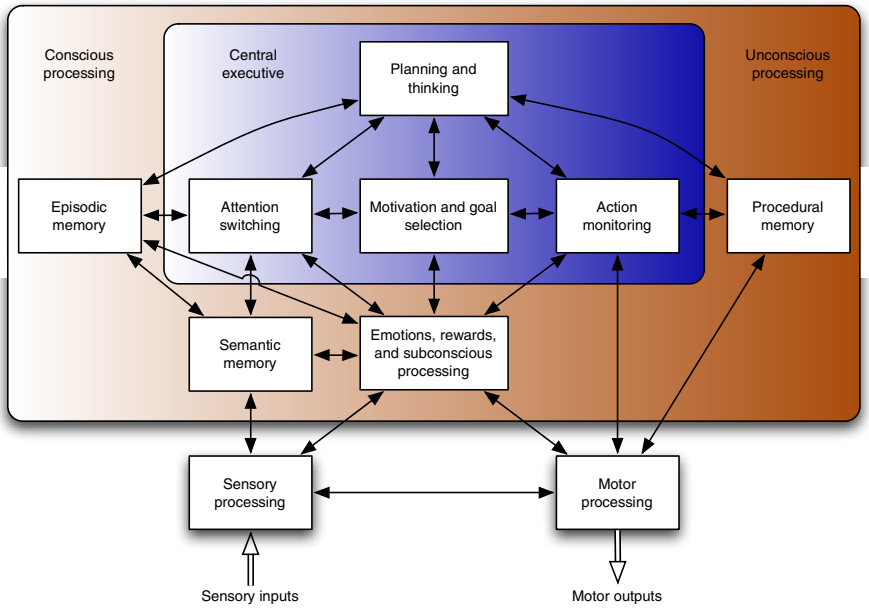


Fig. 1. Computational model of a conscious machine

2.1.1 Sensory-Motor

Figure 1 is divided into three main "areas", the sensory-motor area (outer white region), the episodic memory and learning area (outer orange region), and the central executive area (inner blue region). These blocks could also be referred to as the knowledge, memory, and control blocks. The knowledge block contains inherent and learned skills, activates semantic memories, and performs symbol grounding and reactive motor control. Because skills, reflexes and other abilities are tied integrally into both sensory and motor functions, this block also contains the sensory and motor components. Information comes in via the sensory inputs; it is processed and used to excite the semantic memory, emotional block (in Rewards and subconscious

processing) and motor blocks. Currently our implementation of the Sensory-Motor block is semi-symbolic, in that entire "objects" are taken from the virtual world in which they reside, and associated features are directly presented to the agent. More on this is described in the next subsection.

2.1.2 Memory

Attention switching and memory are closely related to each other, and can be considered interdependent. A portion of attention switching is actually integral to the memory; hence the blocks are not completely separate, although, they are shown that way. This is due to the saccade process. In other words, when the command to saccade is received from the planning block, the memory itself can perform an "attention switch," especially if it's a visual saccade. There are also attention switching thresholds which can be controlled by the planning block and determine whether the machine allows interruptions from the environment. There would be several "levels of attention", such as planning, intended action, and active search. Planning mode is when the machine is blocking out everything but extreme changes in the environment to internally analyze its situation and work out a plan for its current needs. Intended action mode, occurs when the agent is in the process of performing an action, but allows for brief interrupts to evaluate things in the environment. Active search mode is when the agent is actively searching an evaluating the environment. Attention switching becomes an integral part of any realistic agent due to memory and processing limitations. It would be inefficient (and very complex) for the agent to "pay attention" to everything it observes and is present in its memory. Another dependency is that memory stores past experience and thus advises what should be attended to.

The episodic memory, procedural memory, and semantic memory learning capture spatio-temporal sequences of semantic relations. The semantic memory is important for a cognitive agent since it contains contextual knowledge from the input. However, other "reflex" actions can also be excited via sensory input links to the motor and emotional blocks. We conceptualize the organization of the semantic memory as in the SHYNE network presented by Conforth and Meng [10]. SHYNE is a **s**emantic **hy**per **ne**twork capable of knowledge representation and is composed of nodes and links. Nodes represent concepts that can encapsulate or be encapsulated by other concepts. Links represent the relationships between nodes both in terms of "direction" and how closely related they might be.

In terms of implementation, a form of semantic memory [11] is currently being implemented, while leaving the comparatively more difficult episodic memory for later, after the core processes have been implemented.

2.1.3 Central Executive

The Central Executive is the decision making block and determines the system's goals [3]. Once a winner of the internal competition is established, the central executive provides cognitive interpretation of the result, providing top down activation for perception, planning, internal thought or motor functions. It uses

attention switching and mental saccades [9] to streamline the thought processes of the cognitive agent. In this architecture, the machine focuses its attention on the most salient activated region in its working memory, checks its relation to its internal motivations and either performs an action or moves to the next mental saccade, temporarily inhibiting the previously selected concept.

Active perception stimulates representation of the observed scene in the semantic memory. Conscious focus is obtained as a result of attention switching to the most relevant signal that may come from the environment, mental saccades or internal desires. However, conscious thoughts are considered sequentially as they enter the attention focus. There is no competition of unconscious thoughts such as exists in the Baars' Global Workspace Theory model [12] and the selection of a signal to switch attention focus is entirely unconscious. Because of this, the machine is only explicitly aware of the visual objects and/or associations currently under its direct attention. It is not consciously aware of the "background" process and states that compete and direct its attention; rather it examines its environment and associated associations in a sequential (if rapid) manner. Using attention switching greatly simplifies the amount of work the conscious processes need to accomplish, because it only works on a single "focus of attention" at time rather than trying to work on everything simultaneously.

Behind most of this activity is the motivational block, which is continually helping to determine the "importance" of various objects and events and communicating its "needs" to the attention switching and planning blocks. The planning and thinking block takes the current attention focus and motivations into account and attempts to find a solution to the agent's current needs. The last of the blocks in the Central Executive module is the action monitoring block that acts as the interface to the motor functions. It sets tasks to the motor processing and procedural memory blocks, which attempts to carry them out.

Note that there is no clearly defined decision making center. Decisions are actually a result of competition between signals that represent motivations, pains and desires. Furthermore, decisions can be rather fluidic, since competition between the deciding signals can be interrupted by the attention switching signal. As mentioned, the cognitive aspect of the central executive mechanism is predominantly sequential, as a winner of the internal competition is identified and serves as an instantaneous director of the cognitive thought process via the planning block, before it is replaced by another winner through attention switching and changing motivations to act.

2.2 Role of Motivated Learning in the Model

Motivations can change from either internal or external conditions. External conditions are those associated with the sensory input, whereas internal conditions are those associated with a particular plan generated by the planning block or perhaps an emotional context associated with an active episodic memory. Motivations themselves can be either cognitively recognized or unconscious needs. Cognitively recognized motivations would be those that are generated as part of a plan to fulfill some need, whereas motivations arising from unconscious needs are those associated

with more primitive needs and/or feelings and come to machine's focus only after they won attention switching competition.

The ML block is the primary drive behind the system's actions, effectively playing the role of central executive, especially during the early stages of learning. Only later in the learning process will their be enough information in the system for the Planning block to take a more comprehensive role in the system. The ML block receives inputs from the other blocks and calculates the pains or needs associated with the inputs. It is also one of the three affecters of the attention switching block, and can cause a switch in the machine's attention focus. For example, if the machine's primary motivation becomes the resolution of its hunger pain, the motivational block will prompt the attention switching block (in conjunction with the semantic memory and planning block) to find something to resolve its hunger.

Other authors, particularly those who focus on reinforcement learning such as Bakker and Schmidhuber [13] prefer to focus on rewards (i.e. pleasure) rather than "pain". Properly controlled reward systems can work well, however, it has been shown both in simulation and in real-world situations that the use of reward signals can lead to system instability and slow learning in dynamically changing environments. A well known example of this type of problem can be found in Pavlov's work, or in more recent work such as that discussed by Baars and Gage [7] where rats were shown to electrically stimulate their pleasure centers in preference to eating.

3 Simulated Environments

The initial versions of the Motivated Learning software [14] used a very simple simulated environment to demonstrate its characteristics and advantages over RL; however, this is inadequate for testing of more complex behaviors and systems. To overcome the shortcomings of basic predefined environments, we are working on integrating our agent with NeoAxis [15] and other platforms, for example, iCub [16] and Blender [17].

We chose graphically attractive 3D platforms to get both: statistical and visual results. Starting from a rather simple model of a realistic environment we can go on to achieve the simulation of learning close to the complexity of human behavior. Since our cognitive approach is mainly focused on computational abilities we don't need physical body like robots, hence the reliance on more symbolic I/O, but we can treat it as a future perspective and goal. It is very helpful to use computer simulation tools in terms of both cost and effort, because it is expensive to buy or build a physical robot. Current simulators are very flexible and they can easy imitate many real time physical conditions.

3.1 Machine Learning in the iCub Simulator

The iCub cognitive architecture is the result of a detailed design process founded on the developmental psychology and neurophysiology of humans, capturing much of what is known about the neuroscience of action, perception, and cognition. The

architecture itself is realized as a set of YARP [18] executables, connected by YARP ports.

The immediate purpose in developing the cognitive architecture is to create a core software infrastructure for the iCub so that it will be able to exhibit a set of target behaviors for an Empirical Investigations (Looking, Reaching, Reach and Grasp, Reach and Posture, Postural Control in Action, Object Containment, Pointing and Gesturing). See the example in Figure 2.

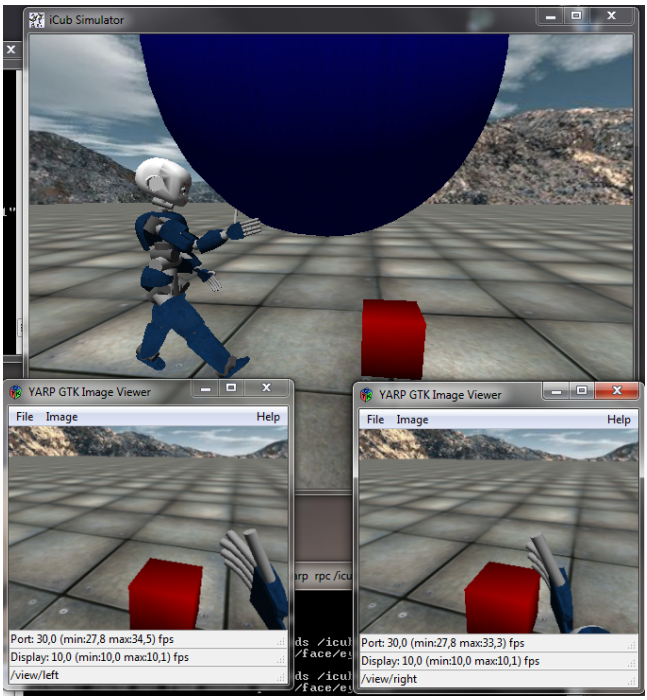


Fig. 2. iCub walking and trying to grasp an object

It is worth remarking that although the iCub cognitive architecture is based on work by Shanahan [19, 20], it is significant that Shanahan's own cognitive architecture does not (yet) incorporate any learning mechanisms.

Currently after some study of machine learning methods and modifications to the iCub simulator source code, a robot can learn to reach and grasp objects. Going forward, we are working on an object recognition and identification algorithms through the iCub vision system to deploy attention and consciousness mechanisms.

Another goal is to improve algorithms used for the iCub's learning process, which are currently based on the Reinforcement learning method [21]. We will try to implement a more efficient method, using a motivated learning mechanism based on goal creation experiment [22].

3.2 Blender

Blender is an open source 3D modeling and rendering application [17] whose main purpose is the creation of computer generated images and animations. Though it is not designed as a tool for simulation, it provides many features that facilitate the development of such an application. There already exists a community of robotics researchers who use Blender [23] for some simulations and there is a drive to improve on this functionality. Blender can be used to aid robotics research in approaches consisting of simple visualization, to simulation and emulation, all the way up to high-level architecture.

The most obvious advantage of using Blender is the high level of graphical detail that can be achieved in real time, as a result of the advanced modeling of meshes, combined with effects such as texturing, lighting and shaders. The visual aspect is important when simulating robotic vision, since the images captured in the virtual world can be realistic enough to be processed with the same algorithms as real images. Blender also offers the capability of using several camera views to follow the evolution of the simulation, displaying a global view of the scenario, as well as views from each of the cameras on–board the various robots. Blender provides the tools necessary to model robots and scenarios with as much detail as required. See the example in Figure 3. Furthermore, Blender also gives immediate access to the Bullet engine for physics simulation. The interface with the modeled objects is already integrated, and the physical properties of objects can be specified in control panels.

3.2.1 Example

Our virtual agent (VEEMA) is placed in an environment together with some resources for example: Farm, ATM, Grocery, Prey, Predator, Pharmacy, Work, and Play [26]. In this simulation model there are three primitive pains: Hunger, Life (avoidance of death, bad health etc), and Boredom. These pains continuously increase with time but they can be reduced i.e. hunger can be reduced by going to Farm, Prey, and Grocery, bad health can be reduced by going to Pharmacy and buying medicine. Our agent has an internal motivational mechanism [24]. The agent, encountering various objects in its environment, will be able to use them; and if it does it will have a direct impact on its condition. For example, the Hunger pain and Life pain lead to the abstract pain of Lack of Money since when the agent buys food or medicine the money available to the agent is reduced. Therefore, the abstract pain of money motivates the agent to earn more money. Going to the ATM and withdrawing money, and so on, reduces the money pain.

The result could be a very flexible and powerful robotics integrated development environment that could become the open source foundation of various tool chains in robotics and other engineering domains (automation, traffic control, gait analysis, motion capturing for animation, etc.) [25]. Although Blender has many advantages it has also some disadvantages, for example: hot-key oriented (not user-friendly in the beginning but finally faster), no reasonable lightmapping solution (reasonable means a lightmap with a different resolution than the original texture), bugs in the Python API (prevents exporting your work), sometimes incompatibility between versions.

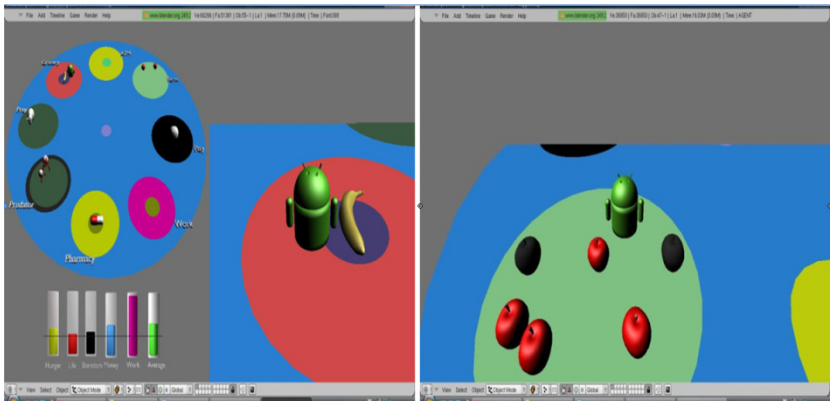


Fig. 3. Search, Find and Learn. The left picture shows the general view illustrating the animated world, the agent and it need signals; the right picture shows a camera view of the agent action.

In our example, a basic learning agent has been achieved. The agent has been able to learn about its environment by learning which actions are useful and which actions are harmful or useless. It has learned to successfully reduce all the pains and keep them close to the threshold value as well as maintain a stable state. Although the agent's functionality is based on simple calculations, we believe that it provides a god starting illustration of what we're trying to accomplish. In Figure 4, the graph presented illustrates the "hunger level" as a function of time. Initially, the pain level goes up to a pre-defined maximum value 100. However, as time progresses, the agent is exposed to the environment and learns how to reduce this pain. So the pain level is brought below the threshold value (set here at 30) by performing useful actions. After the initial series iterations are over and the agent has learned the relationships between its actions and the environment, the hunger level is brought to a stable state and is maintained between the value of 0 and 40.

Fig. 4. Plot of Hunger vs. Time in VEEMA

3.3 NeoAxis

NeoAxis provides a graphical game engine with many existing assets and the ability to add more as needed. It is designed to be easily modifiable by users, and is provided as a free SDK (software development kit) for non-commercial use. The NeoAxis engine itself is based on OGRE (Object-oriented Graphics Rendering Engine) [27].

To embed the ML agent into NeoAxis we decided to modify the game's default AIs and integrate the agent into the decision making part of the code. Additionally, a new class of environmental objects referred to as "Resources" was created to simplify the transition. Integrating the ML agent required providing information from the environment to the agent and receiving and interpreting the agent's responses.

Figure 5 shows a view of the agent (depicted as a woman with a hammer) making it's way toward an object just out of view of the camera. Under normal "play" conditions the red and green lines would not be visible, however, they've been enabled to help visually indicate the agent's actions. In this instance, the agent has decided to investigate a specific resource and is proceeding towards it.

NeoAxis is a relatively powerful platform in that it has access to fully functioning map and resource editors and well as a pair of physics engines." In addition to the creation of custom game resources, the SDK allows programming of specialized behaviors and interactions. In other words, there is a great deal that can be done to create specialized environments from an open ended "moon-rover" scenario to a more closed-world puzzle solving environment.

Fig. 5. In-game Agent

4 Current Work, Future Plans and Conclusions

Current work on the cognitive model presented in this paper is focused on building the model in an "evolutionary" manner. Specifically, we are working at designing and building the core components, then adding additional components as they become appropriate, and adding in complexity to earlier components as needed. For example, our earliest work in motivated learning consisted of the Motivation or Goal Selection block. In this implementation many of the other functions of the cognitive model were rolled into the block and greatly simplified, or left our altogether.

Presently work is being undertaken to expand this implementation by doing several modifications at once. In particular, a semantic memory combined with mental saccading as described in [9] is being implemented. By implementing the attention switching and serializing portions of the machines "thought" process, we are much closer to actual realization of the model of machine consciousness as described in this paper. Initial implementation is expected to be done by using one of the software simulations mentioned in section 3 to provide sensory excitation to the semantic memory. The machine will perform "visual saccades" on the initial memory activations. Then by assessing the winner of the visual saccade, it will perform "mental saccades" on the associations of the winner, which in combination with the needs presented from the motivational components will direct the machine to perform accordingly.

In the paper, we have presented the concept of a computation model of consciousness as a feature of a cognitive agent. We visualize our ideas using three different software 3D architectures: NeoAxis Game Engine, Blender and iCub Simulator. All these platforms are open source, graphically attractive and allow the user to perform effective application development. We can extend the functionality of our agent as much as we want, adding new features and capabilities. Modeling of a realistic environment and agent can go up to achieving emulation of agent learning close to the complexities of human behavior. An additional feature could be to achieve a higher level of complexity integrating specialized work in different areas of artificial intelligence like: speech recognition, robotic vision. With all of this ongoing work we are getting closer to the main goal of designing and implementing an agent capable of successfully negotiating and learning to handle unfamiliar environments.

References

1. Baars, B.: A cognitive theory of consciousness. Cambridge University Press, New York (1988)
2. Aleksander, I.: Impossible Minds: My neurons, My Consciousness. Imperial College Press (1996)
3. Starzyk, J.A., Prasad, D.K.: A Computational Model of Machine Consciousness. International Journal of Machine Consciousness (2011)
4. Rao, A.S., Georgeff, M.P.: BDI Agents: From Theory to Practice. In: Lesser, V. (ed.) Proceedings of the 1st International Conference on Multi-Agent Systems, pp. 312–319. MIT Press (1995)

5. Dastani, M., Dignum, F., Meyer, J.J.: Autonomy, and Agent Deliberation. In: Proceedings of the 1st International Workshop on Computational Autonomy (2003)
6. Wooldridge, M.: Reasoning about Rational Agents. Intelligent Robots and Autonomous Agents. The MIT Press, Cambridge (2000)
7. Baars, B.J., Gage, N.M.: Cognition, Brain, and Consciousness, p. 383. Academic Press (2007)
8. Starzyk, J.A., Graham, J.T., Raif, P., Tan, A.-H.: Motivated Learning for Autonomous Robots Development. Cognitive Science Research 14(1), 10–25 (2012)
9. Starzyk, J.A.: Mental Saccades in Control of Cognitive Process. In: Proceedings of the International Joint Conference on Neural Networks, San Jose, CA (2011)
10. Conforth, M., Meng, Y.: Self-Reorganizing Knowledge Representation for Autonomous Learning in Social Agents. In: Proceedings of the International Joint Conference on Neural Networks, San Jose, CA (2011)
11. Caramazza, A.: How Many Levels of Processing Are There in Lexical Access? Cognitive Neuropsychology 14(1), 177–208 (1997)
12. Baars, B.J.: In the Theater of Consciousness: The Workspace of the Mind. Oxford University Press, New York (1997)
13. Bakker, B., Schmidhuber, J.: Hierarchical Reinforcement Learning Based on Subgoal Discovery and Subpolicy Specialization. In: Proceedings of the 8th Conference on Intelligent Autonomous Systems, Amsterdam, Netherlands, pp. 438–445 (2004)
14. Starzyk, J.A.: Motivation in Embodied Intelligence. In: Frontiers in Robotics, Automation and Control, pp. 83–110. I-Tech Education and Publishing (2008)
15. NeoAxis – all-purpose, modern 3D graphics engine for 3D simulations, visualizations and games, http://www.neoaxis.com/
16. iCub, RobotCub – An Open Framework for Research in Embodied Cognition (2004), http://www.robotcub.org/
17. Blender, http://www.blender.org/
18. YARP - Yet Another Robot Platform, http://eris.liralab.it/yarp/
19. Shanahan, M.P.: A cognitive architecture that combines internal simulation with a global workspace. Consciousness and Cognition 15, 433–449 (2006)
20. Shanahan, M.P.: Emotion, and imagination: A brain-inspired architecture for cognitive robotics. In: Proceedings of the AISB Symposium on Next Generation Approaches to Machine Consciousness, pp. 26–35 (2005)
21. Tamosiunaite, M., Asfour, T., Wörgötter, F.: Learning to reach by reinforcement learning using a receptive field based function approximation approach with continuous actions. Biological Cybernetics 100(3), 249–260 (2009)
22. Raif, P., Starzyk, J.A.: Motivated Learning In Autonomous Systems. In: Proceedings of the International Joint Conference on Neural Networks, San Jose, CA (2011)
23. Community:Science/Robotics, http://wiki.blender.org/index.php/Community:Science/Robotics
24. Starzyk, J.A.: Motivated Learning for Computational Intelligence. In: Igelnik, B. (ed.) Computational Modeling and Simulation of Intellect: Current State and Future Perspectives. IGI Publishing (2011)
25. Blender For Robotics, http://wiki.blender.org/index.php/Community:Science/Robotics
26. Yadav, R.: Design and Simulation of Virtual Environment with Embodied Motivated Agent (VEEMA), MS Project Report, Ohio University (2011)
27. OGRE – Open Source 3D Graphics Engine, http://www.ogre3d.org

Are Pointing Gestures Induced by Communicative Intention?

Ewa Jarmolowicz-Nowikow

Institute of Linguistics AMU Poland
Center for Speech and Language Processing AMU Poland
ewa@jarmolowicz.art.pl

Abstract. The aim of the paper is to present some ideas and observations on the communicative intentionality of pointing gestures. The material under study consists of twenty "origami" dialogue tasks, half of them recorded in mutual visibility (MV) and half in lack visibility (LV) condition. Two participants took part in each dialogue sessions: Instructor Giver (IG) and Instructor Follower (IF). The analysis is focused on selected features of pointing gestures as well as on the semantic aspects of the verbal expressions realised concurrently with pointings: semantic content[1] of verbal expressions realised concurrently with pointing gestures, preceding context of the utterances, place of pointing gestures' realisation in gesture space and spatial perspective of their realisation are taken into consideration as potential cues of communicative intentions

Keywords: pointing gesture, communicative intention, dialogue.

1 Introduction

The issue of communicative intentions behind gestures belongs to the most fundamental problems of multimodal communication studies. Researchers represent diversified opinions concerning the problem. Some of them [1-3 an many others] present the evidence that gestures are intentionally used to communicate. But others [4-6] take the opposite view, arguing that gestures play a role in communication but not as a mean of communicating intentions.

Even though the researchers' views on the communicative intentionality of gestures differ a lot, most of them agree that pointing gestures are a peculiar category of gestures because they are performed, as a rule, with the intention to communicate. Melinger et al. [7] state that "The intensity of the debate about the communicative functions of gestures varies greatly for different types of gestures. Most researches agree that deictic or pointing gestures, which identify real or abstract entities or locations in space are often intended to communicate". According to Masataka [8], the underlying intention of pointings is always to draw someone else's attention to an object or an event of interest. Krauss [6] states that deictic gestures as well

[1] "Semantic content" is here understood as information that contributes to the intended meaning of an utterance [4], [11], [12].

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 377–389, 2012.

as emblematic gestures are as a rule both communicatively intended and communicatively effective. Levelt et al. [9] attribute an important communicative function to pointing gestures by stating that they make deictic utterances complete.

The study conducted by Jarmolowicz-Nowikow and Karpinski [10] provided evidence that not all pointings are driven by communicative intentions. Some of them seem to be performed with only little or no attention directed to the addressee. It was assumed that there are some categories of behaviour of dialogue participants that could indicate communicative intentionality of pointing gestures in the context of the task they were trying to accomplish: the location of gesture in the gestural space, the duration as well as the moment of termination of Pointing Hold Phase2 in pointing gesture's structure and the mutual gaze of dialogue participants. The results of analysis regarding the location of gesture in the gestural space and the duration of Pointing Holds may suggest that they are somehow related to intentionality. However, when some potential nonverbal cues like gaze direction and hold termination were also taken into account as components of more complex communicative structures, the results of the study were less obvious and quite difficult to interpret.

2 The Aim of the Study

It seems reasonable to assume that people have intentions that are externalized in their behaviour. But the analysis of communicative intentionality of gestures is difficult and may be loaded with mistake unless one has a direct access to human mind. As it is presently impossible, the results of such studies may only give some clues or allow to form more precise hypotheses.

The aim of the present study is to make an attempt to determine whether, in some communicative situations, pointing gestures are always performed with intention to communicate. The analysis shows that the realization of pointing gestures is not always receiver-oriented. In the text, the situations of potentially communicative intentional and non-intentional usage of pointing gestures are presented and discussed. Some instances of verbal and nonverbal behaviour are argued to provide clues in the study of intentionality.

The communicative intention is understood as a mental cause for a communicative action intended to be somehow noticed by the addressee [11-14]. It is externalized in gesturing for another to influence her/him (i.e., to change her/his mental state). It is assumed that gestures driven by communicative intention do not necessarily have to be communicative, that is perceived nor interpreted by receiver [10],[15],[16].

The following aspects of communicational behaviour were taken into consideration in order to distinguish potential cues of communicative intention behind pointing gestures:

[2] Pointing Hold Phase - a phase in a structure of pointing gesture phrase that is an equivalent to Post-stroke Phase in Kendon's gesture phrase. A modification to the description of gestural phrase structure for pointing gestures was proposed by [10].

- the place of pointing gesture realization in gestural space;
- relative to IG or IF perspective of pointing gesture realization;
- the semantic content of verbal expressions realised concurrently with pointings;
- the preceding context of the analysed verbal expressions.

The postulate of economy and postulate of integration of gesture and speech were taken under consideration analysing material studied.

- The postulate of communication economy [11], [17-19] according to which the behaviour in communication should be oriented toward the minimum effort that is necessary to achieve the intended result. The concept of economy in communication refers traditionally to spoken language, however it may be extended to analysis of multimodal communication taking gestures into consideration;
- The postulate of the integration of gesture and speech [20],[21] according to which gesture and speech are forming an integrated system. "Speech and synchronous gestures form a tightly bound unit, capable of resisting outside forces attempting to separate them" [21]. It is possible to distinguish two kinds of relations between pointing gestures and speech: 1. when pointings carry *primary, informationally foregrounded information* and speech is *supportive* 2. when pointings carry *secondary, informationally backgrounded information* and speech is *central* [22].

3 Material under Study

The material under study consists of 20 sessions of "origami" dialogue task from the DiaGest2 project (e.g., [23],[24]) recorded in two conditions: ten in the mutual visibility condition (MV; subject could see each others) and ten in the lack of mutual visibility condition (LV). The dialogue task in the MV as well as the LV condition involved the reconstruction of a paper figure which was visible to one person (Instruction Giver, IG) and not visible to the second (Instruction Follower, IF). In the MV condition the paper figure was on IG's desk and was screened so IF could not see it. In the LV condition, a screen was situated between IG and IF so they could not see each other. IF was provided with all the necessary materials to construct her/his “copy” of the original figure. The duration of the sessions was limited to 5 minutes. The distance between the participants in the MV condition was approximately 3 meters. Seventy-eight pointing gestures and verbal expressions realised concurrently were analysed in the MV and 50 in the LV condition. The sessions were recorded using four (MV) or two camcorders (LV) and independent high-quality audio recording (Fig. 1 and 2). The subjects were polish university students. Video recordings were converted to MPEG1 format and synchronised in ELAN (by MPI). Gesture annotation was carried out in ELAN while orthographic transcriptions, prosodic annotations and some other tagging were done in Praat [25] and imported into ELAN.

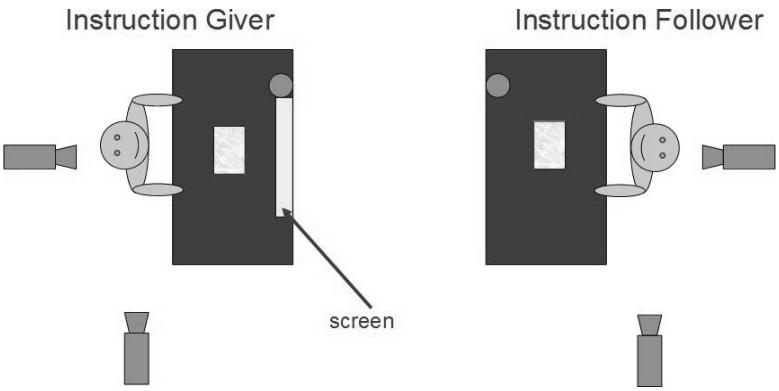


Fig. 1. Recording settings - mutual visibility condition (MV)

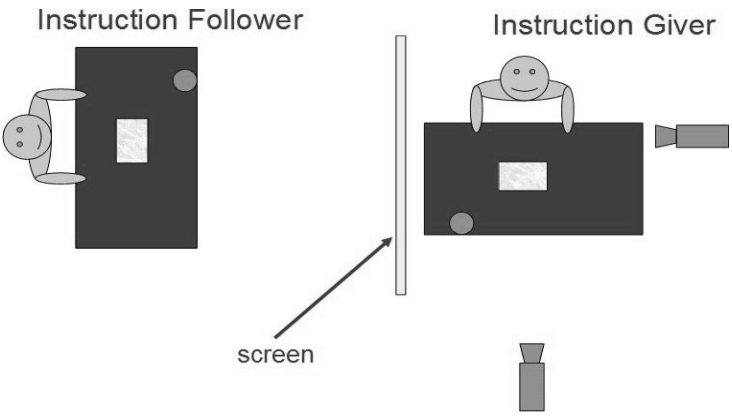


Fig. 2. Recording settings - limited mutual visibility condition (LV)

4 Verbal and Nonverbal Cues of Potential Intentionality of Pointing Gestures

4.1 Place of Realization in the Gestural Space

Pointing gestures were identified on the basis of their form - a hand directed toward an object, location or direction. Pointing gestures tagged in the studied material are performed by hand movements, however they are not restricted to a specific form or shape of the movement trajectory. As the communicative intention is to be seen and interpreted it is assumed that pointing gestures with intention to communicate are produced in the gestural space visible to IF.

In order to describe the distribution of IG's pointing gestures in the gestural space, two dimensions were taken into account:

- the horizontal position of the palm in relation to the figure (*behind, over, in front of, extremely in front of*);
- the vertical position of the palm in relation to the IG (*above head, head, shoulder, chest, abdomen, belt, below belt*).

The location of the palm in the peak of each pointing gesture (i.e., at the end of the Pointing Movement3) was described in the above dimensions for both conditions (MV and LV).

The analysis of communicative intentionality carried out by Jarmolowicz-Nowikow and Karpinski showed that in the MV condition, pointing gestures were most often realised on the level of *shoulders* and *chest;* no pointings occurred on *below belt* level. In the LV condition no pointings were realised on *shoulder* and *chest* level, while most of them were produced on *belt* and *below belt* level [10]. The place of gesture realisation in MV might have been a sign of communicative intention behind pointing gestures as the original figure on IG's desk was hidden behind a small screen so the IF could not see it. Therefore, the realisation of gesture with intention to communicate required producing gestures in the location visible to IF. In the analysis carried out for the present paper, the place of pointing gestures realisation (visible or not visible to IF) is regarded as a key indication of communicative intentionality.

4.2 Perspective Relative to IF or IG of Pointing Gestures Realisation

Another nonverbal cue taken into consideration is the perspective (relative to IF or IG) of the realisation of pointing gesture. In order to analyse this aspect of gesture, the following conditions were taken into account:

- IG points to his/her paper figure or IF's paper figure (or components of the figure under construction);
- IG points to a direction from his/her point of view or from IF's perspective (e.g. IG says: *to the right* and indicates IG's right side which is left to the IF staying opposite to IG).

The analysis of pointing gestures in the LV condition shows that all of them are realised from IG's point of view. It is absolutely natural that IG having no visual access to IF's figure, points at his/her figure and indicates direction form IG's viewpoint. In the MV condition, pointing gestures are realised from both perspectives. It is assumed that pointing gestures driven by communicative intentionse are produced with respect to IF's spatial co-ordinates to help IF to follow next steps of the dialogue task.

---

[3] Pointing Movement Phase (PM) - a phase in a structure of pointing gesture phrase that is an equivalent to stroke in Kendon's gesture phrase. A modification to the description of gestural phrase structure for pointing gestures was proposed by [10].

4.3 Semantic Content of Verbal Expressions Realised Concurrently with Pointing Gestures

On the basis of the postulate of the economy and effort optimalisation [11], [17-19] and the postulate of gesture and speech integration [20], [21], it is assumed that gesture can compensate for the information omitted in speech. Therefore, it is hypothesised that the less informative the verbal expression referring to object, location or direction in the MV condition is, the more probability that pointing gestures are produced with the intention to communicate. The following steps were undertaken in order to check whether there is an interdependence between the semantic content of the verbal expressions taken under consideration and the potential communicative intentionality of pointing gestures.

Verbal expressions realised concurrently[4] with pointing gestures were divided into three groups:

- referring to an object (answering the question *what*? e.g. *Ten następny narożnik* = ″This next corner");
- referring to a location (answering the question *where*? e.g. *Na mniejszej ścianie* = "On the shorter wall");
- referring to a direction (answering the question *which direction*? e.g. *W lewo* = ″To the left").

The division of the utterances into three groups is compatible with the categorisation of the typical function of pointing gestures: indicating objects, locations and directions [26]. Semantic analysis of verbal expressions concurrent to pointing gestures is focused on the elements referring directly to gesture e.g. in the expression *Właśnie widzisz, w tych rogach* (So, you can see, in these corners) only the expression *w tych rogach* (in these corners) was taken into consideration.

The next stage of the analysis was to divide all the expressions into *not semantically autonomous* (e,g, *this one* / *there*) and s*emantically autonomous*5 (e.g. *this upper right corner*). For the purpose of classification, the utterances were analysed in terms of:

- the number of semantic elements constituting verbal expressions;
- the semantic definiteness of components constituting verbal expressions;
- the reference of semantic elements (to a unique part of the figure or to one of the identical parts of the figure e.g. corner in a square figure).

It is assumed that the less semantic elements constitute verbal expressions and the less definite elements are, the more probability that the expression is context-dependent and thus not fully comprehensible for IF. If the verbal expression contains the only one semantic element that refers to the one of the identical parts of the figure (e.g. corner of the square) it is regarded as context-dependent.

---

[4] The concurrent realisation of verbal expression and pointing gesture is defined as a situation when the verbal expression is realised while pointing gesture is being produced.

[5] The term *semantically autonomous* is relative as any expression is not fully autonomous. More abut the problem in section 4.4.

4.4 The Preceding Context to Analysed Verbal Expressions

The context preceding the expression realised concurrently to pointing gestures is assumed to be one of the components that should be taken under consideration while interpreting pointing gesture as intentional communicatively. The realisation of pointing gesture (1) in the area of the gesture space visible to IF, (2) from IF's perspective, (3) with semantic content of expression concurrent to pointing gesture not sufficient for IF to follow IG's instruction might suggest that pointing gesture is potentially communicatively intentional. However, the information provided by gesture might have already been provided by the preceding context. The realisation of pointing gesture (1) in the area of the gesture space visible to IF, (2) from IF's perspective and concurrently (3) with a semantically autonomous verbal expression does not reduce probability that gesture is produced with intention to communicate, however the communicative function is less important - IF does not have to pay attention to gesture.

One of the problems related to the incorporation of the preceding context into analysis is that it is difficult to assess how far back the analysis should reach. In fact, the entire previous dialogue may contain important contextual information. In practise, it seems necessary to decide arbitrarily about the range of the context taken into consideration or to define some applicable measures or formulae for its determination. Otherwise, the interpretation of the effects of context may remain an open-ended investigation of people's knowledge of the commonsense world [27]. It was decided that a piece of previous dialogue between IG and IF that directly refers (or is focused on) the part of the object, the location or the direction that verbal expression concurrent to pointing refers to will be taken into account. The dialogue task focused on instructing how to build a paper figure consists of certain clearly distinguishable stages. It it possible to determine the boundaries of these stages where the participants start to concentrate on a certain part of the figure or location.

The utterances under study are divided in the two categories:

- PC (preceding context) when the meaning of the expression concurrent with pointing gesture is largely dependent on the preceding context;
- NPC (no preceding context) when the meaning of the expression concurrent with pointing gesture is largely independent from the preceding context.

5 Observations and Remarks on the Communicative Intention behind Pointing Gestures

5.1 Pointing Gestures with Potential Communicative Intention in the MV Condition

The category of pointing gestures that seem to express the information necessary to continue dialogue task but omitted from speech was distinguished in the analysed material in the MV condition. These gestures may be regarded as realised with potential intention to communicate. The following situations may be considered symptomatic for the pointing gestures classified as communicatively intentional in the present context:

- a pointing gesture is realised in gestural space visible to IF;
- a pointing gesture is realised from IF's point of view;
- the verbal expression concurrent with a pointing gesture is not semantically autonomous;
- the verbal expression concurrent with a pointing gesture is semantically largely independent from the preceding context.

It is important to emphasize that when the verbal expression uttered concurrently with a pointing gesture is semantically autonomous or depends on the preceding context it does not necessarily mean that the pointing gesture is not realised with intention to communicate. In such a communicative situation, pointing gestures realised in the visible area of the gesture space and taking the IF's spatial perspective under consideration may cause that the communicative intentionality of the behaviour is even more noticeable [16].

Table 1. The examples of situations where pointing gesture is produced with potential intention to communicate (semantic content of the verbal expression is not semantically autonomous; previous context does not provide relevant information)

Verbal expression	Spatial perspective	The place of realisation in gestural space
tutaj [here]	IF's desk	shoulder/in front of
w tamtem narożnik [to that corner]	IF's desk	shoulder/extremely in front of
na tamtym boku, [on that side]	IF's desk	shoulder/extremely in front of

Table 2. The examples of situation where pointing gesture is produced with a potential intention to communicate (the meaning of the verbal expression is semantically autonomous or dependent on preceding context)

Verbal expression	Spatial perspective	The place of realisation in gestural space	Context
w tym prawym, bliżej mnie rogu [in that right, closer to me corner]	IF's desk	chest/over	expression semantically autonomous
ta przy tobie [this near you]	IF's desk	shoulder/over	PC
tam (...) w tych rogach [there (…) in these corners]	IF's desk	chest/ extremely in front of	PC

5.2 Pointing Gestures with No Communicative Intention in the MV Condition

The group of pointing gestures that are probably not driven by IG's communicative intentions, thus not IF-oriented, is also distinguished in the analysed material. These pointing gestures are probably produced without intention to be noticed nor interpreted by IF.

The pointing gesture may be interpreted as produced with no intention to communicate when the following conditions are fulfilled:

- pointing gesture is realised in the area of the gestural space that is invisible to IF;
- pointing gesture is realised from the IG's perspective;
- the verbal expression concurrent with pointing gesture is classified as semantically autonomous or as not semantically autonomous but it is dependent on preceding context.

Table 3. The examples of situations where pointing gestures are produced with no intention to communicate (the meaning of the verbal expression is not semantically autonomous and depends on preceding context)

Verbal expression	Spatial perspective	The place of realisation in gestural space
tego [this]	IG's figure	belt/over
prawy róg [right corner]	IG's figure	abdomen/behind
tamta [that]	IG's figure	abdomen/over

In the examples presented above, communicative intentions are realised mainly by the verbal aspect of utterance. Even though the verbal expression concurrent to pointing gesture is not autonomous, it is strongly dependent on the preceding context. Pointing gestures are realised on the level of belt or abdomen, probably not visible to IF (because of the presence of the screen that hides the figure on the IG's desk). However, it was noticed that pointing gestures realised in the part of the gestural space not visible to IF might also co-occur with the verbal expressions with semantic content not sufficient for IF to follow next steps of the dialogue task and with no preceding context. The analysis of the material suggests that this kind of situation was not infrequent in the MV condition. It is assumed that such a situation takes place in the MV condition as IG is aware that mutual visibility gives him/her possibility to track IF's activity. The IG probably does not pay much attention to the informative precision of verbal as well as nonverbal aspects of utterance as the possibility of tracking the IF's behaviour gives the IG a chance to formulate subsequent utterances and realise gestures more precisely. Such a situation almost does not take place in the LV condition. The analysis of verbal expressions concurrent with the pointing gestures referring to object in the LV condition shows that all of the analysed verbal expressions were semantically autonomous or dependent on the preceding context.

Table 4. The examples of situations where pointing gesture is most probably produced with no intention to communicate (the content of the verbal expression is not autonomous, the preceding context does not provide necessary information)

Verbal expression	Spatial perspective	The place of realisation in gestural space
ta [this] ItemDet	IG's figure	abdomen/behind
lewy [left] ItemDet	IG's figure	abdomen/behind

Table 5. The examples of contradiction between the semantic content of a verbal expression and the pointing gesture (the content of the verbal expression is autonomous; no relevant information provided by the preceding context)

Verbal expression	Spatial perspective	The place of realisation in gestural space
twoją prawą [your right]	IG's perspective	behind / shoulder
bliżej ciebie [closer to you]	IG's perspective	chest/behind

Another observation made while analysing origami recordings concerns the contradiction between the semantic content of expressions referring to a direction and the accompanying pointing gestures. Some of the verbal expressions explicitly describe directions from the IF's viewpoint (e.g. *your left*). A few such situations were observed where pointing gesture indicating directions and produced in the visible section of the gesture space visible to IF was realised from IG's perspective.

In the examples above, pointing gestures indicate direction from IG's perspective (e.g. the right hand side from IG's perspective which is the left hand side from IF's perspective). In such a context, pointing gestures may be supposed to lack communicative intentionality.

6 Discussion

It is not difficult to understand that a given type of action (looking through the window) may be driven or not driven by intentions [28]. The difficulties consist in the interpretation when action is driven by certain intentions. Pointing gestures are nonverbal activities that are regarded by most of the researchers as realised with communicative intentions [6-9 and many others]. The reason for such a state of affairs may be twofold. The interest of pointing gestures concerns mainly the development of communication skills in preverbal children, when pointing gestures substitute speech and are used to signal children's intentions [29,30]. Another reason for accepting communicative intentionality of pointing gestures might be a narrow understanding of

their definition. It seems that most of the examples of pointing gestures adduced in the literature to prove their communicative intentionality focus on the gestures based on the extension of the index finger realised concurrently with the expression with indefinite pronoun (e.g., *this*, *that*), where the function of pointings is evident. However, the use of pointing gestures as well as the forms they take is much more differentiated.

In the present study, it was assumed that pointing gestures are not induced by communicative intention 1) when they are realised in the gestural space not visible to IF and 2) when a contradiction between the verbal expression referring to direction and the pointing gesture indicating direction takes place. A significant number of such situations was found in the material under study.

According to Krauss [6], it is reasonable to regard the gesture as intended to aid the receiver. However, the results of his experiment show that most of the participants requested to show the "clockwise direction" to the interlocutor did it from their own perspective which was defective from the addressee's point of view. Some of the situations show that indicating direction by pointing may facilitate IG's mental representation of the relevant spatial relations, and thus not to communicate intentionally. It refers to the situation when complex spatial relations are described and pointing is realised in the section of the gesture space visible to IF but produced from IG's point of view. However, in the "origami" recordings some situations were found where IG held his/her hand in the area of the gesture space visible to IF and s/he was "negotiating" the point of reference for indicated direction (*e.g. so, my right is also your right, isn't it*?). It was noticed that despite "negotiating", the direction was indicated from IG's point of view. Such a situation may suggest that indicating direction might be confusing for IGs thus pointing at wrong direction does not have to be a symptom of the lack of communicative intention. It may be assumed that in such situation IG intends to perform the pointing gesture for IF, however the difficulties with description of the spatial relations cause unintended mistakes.

The studies on the communicative intentionality of pointing gestures based on the behavioural methodology can be hardly unequivocal and satisfactory. The phenomenon of “intentionality” itself has been discussed by philosophers and neuroscientists alike but its understanding is still not clear. However, small steps forward can and should be made to gradually bring researchers towards a better understanding of the problem.

Drawing someone else's attention to an object or event is not the only one function of pointings. The observations made in the present study may suggest that most of the pointing gestures have communicative function but not all of them are produced with intention to communicate. All pointing gestures support the process of communication to a certain extent (some facilitate speech production [4] and thereby perform a communicative function) but only some of them are communicatively intentional.

The studies on intentionality described in the present text will be extended in future with the quantitative analysis of gestural and verbal behaviour that would allow for the statistical testing of some hypotheses on the frequency and context of intentional and non-intentional pointing gestures in the “origami” recordings.

References

1. Bavelas, J., Kenwood, C., Johnson, T.: An experimental study of when and how speakers use gestures to communicate. Gesture 2, 1–17 (2002)
2. Özyürek, A.: Do speakers design their co-speech gestures for their addressess? The effects of addressee location on representational gestures. Journal of Memory and Language 46, 688–704 (2002)
3. Kendon, A.: Do gestures communicate? A review. Research on Language and Social Interaction 27(3), 175–200 (1994)
4. Krauss, R.M., Chen, Y., Chawla, P.: Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In: Zanna, M. (ed.) Advances in Experimental Social Psychology, pp. 389–450. Academic Press, San Diego (1996)
5. Morrel-Samuels, P., Krauss, R.M.: Word familiarity predicts temporal asynchrony of hand gestures and speech. Journal of Experimental Psychology: Learning, Memory, and Cognition 18(3), 615–622 (1992)
6. Krauss, R.M., Hadar, U.: The role of speech-related arm/hand gestures in word retrieval. In: Campbell, R., Messing, L. (eds.) Gesture, Speech, and Sign, pp. 93–116. Oxford University Press, Oxford (1999)
7. Melinger, A., Levelt, W.J.M.: Gesture and the communicative intention of the speaker. Gesture 4(2), 119–141 (2004)
8. Masataka, N.: From index-finger extension to index-finger pointing: Ontogenesis of pointing in preverbal infants. In: Kita, S. (ed.) Pointing. Where Language, Culture, and Cognition Meet, pp. 68–85. Lawrence Erlbaum Associates, Mahwah (2003)
9. Levelt, W.J.M., Richardson, G., Heij, W.L.: Pointing and voicing in deictic expressions. Journal of Memory and Language 24, 133–164 (1985)
10. Jarmolowicz-Nowikow, E., Karpinski, M.: Communicative Intentions behind Pointing Gestures in Task-oriented Dialogues. In: Proceedings of GESPIN 2011, Bielefeld (2011)
11. Grice, H.P.: Utterer's Meaning and Intentions. Philosophical Review 78, 147–177 (1969)
12. Searle, J.R.: Speech acts: An essay in the philosophy of language. Cambridge University Press, Cambridge (1969)
13. Recanati, F.: On Defining Communicative Intentions. Mind and Language 1, 213–242 (1986)
14. Bach, K.: On Communicative Intentions: A Reply to Recanati. Mind and Language 2(2), 141–154 (1987)
15. Taillard, M.O.: Beyond Communicative Intention. UCL Working Papers in Linguistics 14, 189–207 (2002)
16. Jaszczolt, K.M.: Default Semantics. Foundations of a Compositional Theory of Acts of Communication. University Press, Oxford (2005)
17. Horn, L.: Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In: Schiffrin, D. (ed.) Meaning, Form and Use in Context (GURT 1984), pp. 11–42. Washington Georgetown Press (1984)
18. Carston, R.: Relevance theory and neo-Griceans: A response to Lawrence Horn's 'Current issues in neo-Gricean pragmatics'. Intercultural Pragmatics 2(9), 303–319 (2005)
19. Vicentini, A.: The Economy pronciple in Language. Notes and Observations from Early Modern English Grammars. Mots, Palabras, Words 3, 37–57 (2003)
20. Goldin-Meadow, S.: How our Hands Help us Think, Cambridge, Massachusetts, and London, England (2005)
21. McNeill, D.: Gesture&Thought. University of Chicago Press, Chicago (2005)

22. Enfield, N.J., Kita, S., de Ruiter, J.P.: Primary and secondary pragmatic functions of pointing gestures. Journal of Pragmatics 39(10), 1722–1741 (2007)
23. Szczyszek, M., Karpiński, M.: Qualitative and quantitative analysis of lexical, quasi- and non-lexical components of utterances in task-oriented "origami" dialogues". Investigationes Linguisticae XXII (2011)
24. Karpinski, M., Jarmolowicz-Nowikow, E.: Prosodic and Gestural Features of Phrase-internal Disfluencies in Polish Spontaneous Utterances. In: Proceedings of Speech Prosody 2010 Conference, Chicago (2010)
25. Boersma, P., Weenink, D.: Praat: Doing Phonetics by Computer. A computer programme (ver.5.1) (2010)
26. Kita, S.: Pointing. A Foundational Building Block of Human Communication. In: Kita, S. (ed.) Pointing. Where Language, Culture, and Cognition Meet, pp. 1–8. Lawrence Erlbaum Associates, Mahwah (2003)
27. Hobbs, J.R.: Information, Intention, and Structure in Discourse. In: Preceedings of the NATO Workshop on Burning Issues in Discourse, Maratea, pp. 41–66 (1993)
28. Norris, S.: Analysing multimodal interaction. A methodological framework. Routledge Taylor and Francis Group, New York (2004)
29. Crais, E., Douglas, D.D.: The Intersection of the Development of Gestures and Intentionality. Journal of Speech, Language, and Hearing Research 47, 678–694 (2004)
30. Iverson, J.M., Goldin-Meadow, S.: Gesture Paves the Way for Language Development. A Psychological Science 16, 367–371 (2005)

TV Interview Participant Profiles from a Multimodal Perspective

Maria Koutsombogera[1,2] and Harris Papageorgiou[1]

[1] Institute for Language & Speech Processing, Artemidos 6&Epidavrou, 15125 Athens, Greece
[2] University of Athens, Department of Linguistics, University Campus, 15784 Athens, Greece
{mkouts,xaris}@ilsp.gr

Abstract. The study presented in this paper attempts to provide evidence about prominent features in the interactional behavior of TV interview participants that could lead to testable predictions about their communicative profiles. Based on a multimodally annotated interview corpus, we explore the behavior of two groups representing the two basic discursive roles of this domain, namely the interviewers and the interviewees. We describe the profile of the speakers as outlined by their non-verbal activity, in terms of preferred modalities employed as well as the specific goals and functions that they aim to accomplish through the use of non-verbal expressions (gestures, facial expressions, torso movements). From this perspective, we aim to discover possible patterns of non-verbal behavior that are in line with the communicative actions that each role is supposed to perform, as well as significant differences or distinctive features attested in the behavior of each group.

Keywords: multimodal corpus, TV interviews, non-verbal expressions, contextual analysis, communicative roles, multimodal interaction.

1 Introduction

The analysis of prominent communicative features in face-to-face interaction can profit from corpus-based descriptions of targeted phenomena, due to the rich set of instances attested in a situated context as depicted in an annotated resource. Face-to-face interaction involves the employment of distinct modalities, all contributing to the communication of the speakers' message. The communicative significance of gestures, facial expressions and body posture has been studied in terms of transferring information, communicating emotions or coordinating the interaction [1, 2, 3]. The decoding of information conveyed in all modalities is important to both understanding and generation of language, to the description of the detailed ways in which the linguistic system and the gestural system interact, as well as to the communicative goals and the communicative profiles of the speakers. Evidence of this communicative behavior originating from natural or partially controlled contexts unveils a wide range of social and interactional parameters that are crucial for the interpretation of the non-verbal behavior.

In this paper we aim to explore whether there is a generic profile of communicative interactional behavior for the roles that the speakers assume in the television

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 390–397, 2012.

interview media genre. In order to answer this question, and given the importance of the context in shaping the discursive properties in interviews - and subsequently the production of the participants' multimodal behavior - we study the effect of the context in the sense of (a) the different interview types and (b) the communicative roles that appear in them.

Specifically, we aim to describe the profiles of the two discursive roles of a television face-to-face interview, namely the role of the host (interviewer) and the role of the guest (interviewee) based on their non-verbal behavior. In this context, by non-verbal expressions we mean the modalities of gestures, facial expressions and torso movements employed in the flow of the discussion. Most importantly, each instantiation of a non-verbal expression is accompanied by a set of interactional functions that it performs, in terms of turn management, feedback responses and expression of emotions and attitudes. In this perspective, what we think of as a *profile* of each role is the set outlined by the preferred values of the aforementioned features.

This study was carried out by exploiting an annotated corpus of 12 TV interviews. The non-verbal behavior of the participants was studied in the overall corpus as well as in the two subcorpora in which it consists, namely a political and a cultural one. From this perspective, we aim to discover possible patterns of non-verbal behavior that are in line with the communicative actions that each role is supposed to perform, as well as significant differences between the distinct datasets. The findings are collected through a comparison of the frequency and the values of the features at hand (a) between the two speaker roles within the same dataset and (b) within the same speaker role across the different datasets.

Our approach is corpus-based, and it has the goal of exploring the communicative context which raises respective discourse and communicative demands. In general, contexts may vary and they trigger a multitude of communicative needs that the speakers aim to fulfill. The planning of their discourse includes the employment of non-verbal expressions which implement their intentions and perform interactional functions. Thus, the speakers use all their expressive modalities as a flexible resource to meet the demands of a given communicative context; subsequently, the study of the various contexts and the respective pragmatic uses may reveal significant findings interpreting the choices the speakers make with regards to their interactional behavior. A relevant study exploring the varying degrees of institutionality as attested in three interviews [4] has shown that there are indeed differences in the resources that the speakers employ in line with the communicative demands. Therefore, all non-verbal attestations that have been annotated in the corpus have a communicative dimension, either in terms of intent on behalf of the speaker or as part of the impact on the recipient and the flow of the interaction itself.

2 Corpus

The interview corpus examined is of approximately 170 minutes duration and it consists of 12 interviews, 7 of them pertaining to the political genre (6 hosts – 7 guests), and 5 of them to the cultural genre (5 hosts – 5 guests). In this study, genre is

defined on the basis of the guest identity as well as the topic under discussion. For example, the guests of political interviews are usually parliament representatives talking about current political issues (government measures, elections etc.), while in the cultural ones, the guests are actors, writers, musicians and TV personas presenting their activities but also talking about personal issues. The interviews follow the typical structure of a 2-person discussion program, i.e. the host addresses questions to the guest and the discussion follows an agenda which is predefined in either a strict or less controlled manner. During the interview, the participants present their argumentation to the discussion topics and they elaborate on their opinions. In this context, they produce a multitude of non-verbal expressions co-uttered with speech.

In each interview, the audio signal has been transcribed[1] and the output has been further segmented and annotated in terms of dialogue acts[2] [5]. The video annotation[3] deals with the labeling of the non-verbal expressions (facial, hand and torso movements) co-occurring with speech, at multiple levels: (a) identification and marking on the time axis, and assignment of (b) respective semiotic type, (c) turn management type (d) feedback type, (e) attitudes and emotions expressed [6] and (f) semantic relations with speech. The levels and labels used in the annotation scheme are mainly inspired by the MUMIN coding scheme notation[4] [7].

Both audio and video signals as well as the annotations are perfectly synchronized; the overall set of annotation levels is distinguished by speaker, and all the annotation metadata are integrated into xml files.

The goal of this corpus building process was to study the interplay between speech and non-verbal expressions in its communicative dimension. It was more than evident, however, that more light needed to be shed on the range of the attested contexts. Interviews pertaining to the political subcorpus feature institutional characteristics such as predefined agenda and discourse strategies [8]. On the other hand, the communicative behavior in the cultural corpus follows the properties of casual settings, i.e. flexibility in terms of agenda and turn predictability and more likely to include casual and unpredictable behavior [9]. Accordingly, the discourse types and dialogue acts employed, to which the interview content could be categorized, conform to the demands of the context, resulting in a variety of types, such as argumentation, information, narration, etc.

3 Data Analysis

The time frame in which speakers appear on screen, and therefore their non-verbal activity is visible, may vary. In certain cases a speaker may hold the floor but does not appear in the video; though the voice is heard, the video may focus on the co-locutor or on a location on the set, etc. Thus, simply counting the instances per speaker and

[1] Transcriber (`http://trans.sourceforge.net/`)

[2] General-purpose functions only.

[3] ELAN (`http://www.lat-mpi.eu/tools/elan/`)

[4] Specifically, all labels attributed to the features of gestures, facial expressions, semiotic types, turn management, feedback (except for emotions/attitudes).

per interview would result in the wrong conclusions, as each speaker holds the floor and appears on screen in different time slots. Due to this fact, all measurements consisted in calculating the ratio of the instances (produced non-verbal expressions) of each speaker group per the total number of seconds that it holds the floor, for the overall corpus and for the 2 subcorpora. Then, an independent sample t-test was used to calculate the statistical significance of the attested differences. The dependent variable is the ratio of each value (instance) for each speaker group per the total number of seconds that it holds the floor. The independent variable is of two conditions: (a) both speaker groups per interview genre and (b) the same speaker group in the two different interview subcorpora.

3.1 Production of Non-verbal Expressions

The frequency of the produced non-verbal expressions is the same in all cases, independently of the genre or the role group. This means that, on an average, a speaker produces 0.9 non-verbal expressions per second. Table 1 below lists the number of non-verbal expressions annotated in the corpus per speaker group and per genre, as well as the duration of turns of the overall corpus, the subcorpora and the speaker groups.

Table 1. Number of non-verbal expressions annotated in the corpus and duration of turns

	Number of non-verbal expressions			Turn length (sec.)		
	all	hosts	guests	all	hosts	guests
Overall corpus	10078	3097	6981	10702	3505	7197
Political corpus	8612	2747	5865	8615	2975	5640
Cultural corpus	1466	350	1116	2087	530	1557

Speakers tend to produce more facial expressions (55%), followed by gestures (35%) and less torso movements (10%). The comparison of speaker groups in each modality shows that (a) there is no significant difference between host and guests, no matter what the modality is, and (b) there is a difference within the same speaker groups, across genres, in the modality of gestures. Production of gestures presents a different rate among the political and the cultural corpus. In fact, the number of gestures in the political corpus is almost double compared to the cultural corpus. A possible explanation would be the fact that the argumentative discourse, which is prevalent in the political interviews, is non-verbally expressed with a multitude of quick and brief rhythmic gestures (beats). Table 2 lists the p-values related to measurements conducted with regards to the produced and annotated non-verbal expressions. On the left column the conditions are described. The first three rows concern the comparison between the two speaker groups per interview genre, political, cultural, or overall corpus (all). The following rows present the comparison of the different corpora within the same speaker group, hosts and guests accordingly.

Table 2. P-values for the comparison of non-verbal expressions between speaker roles and between the two subcorpora

Groups	P-values (t-test)		
	Facial expressions	Gestures	Torso
Host vs. Guest			
all	0.67	0.98	0.95
political	0.60	0.91	0.97
cultural	0.91	0.78	0.48
Host			
all vs. political	0.99	0.77	0.83
all vs. cultural	0.96	0.05	0.33
political vs. cultural	0.96	**0.03**	0.25
Guest			
all vs. political	0.90	0.63	0.87
all vs. cultural	0.66	**0.03**	0.58
political vs. cultural	0.59	**0.02**	0.49

3.2 Subtypes

Each non-verbal expression identified is accompanied by a set of features which describe the actual semantics of the distinct modalities in terms of semiotics and communicative functions, i.e. Semiotic Types, Feedback, Turn Management, Semantic relations and Emotions/attitudes. Below, for each feature, we list only the values of the annotation scheme for which a significant difference has been noted between either the two speaker groups or the two subcorpora (cf. Tables 3 and 4). Values in both tables are explained in the following paragraphs.

Semiotic Types. A significant difference has been observed between the two speaker groups regarding iconic and symbolic values, with the guest group producing more expressions of these types. If we compare the attributed semiotic types among the two genres, we attest that the iconic type is mostly prevalent in the cultural subcorpus. The explanation for this observation probably lies in the fact that iconic gestures are linked to data of narrative type, which are common in the cultural corpus.

Feedback. Guests in political interviews seem to produce more negative feedback (cf. Table 3, value: give/non-accept), i.e. that they do not accept what their interlocutors say, probably as a reaction to certain hostile behavior of the hosts expressed through criticism and adversarial questions.

Turn Management. Turn management values describe the coordination of the turns flow in the course of the interview, as participants take, offer, accept, yield or complete their turns. As expected, the hosts are the ones who usually offer the turn and subsequently the turn is accepted by the guests. What is interesting is that guests in political interviews rarely complete their turn without being interrupted or without being forced by the interlocutors to yield it.

Tables 3 and 4 list only the values of the annotation scheme for which a significant difference has been noted between either the two speaker groups (*Host vs. Guest*) or

Table 3. Statistically significant values in non-verbal expression subtypes between speaker roles and between the two subcorpora

	Semiotic types	Feedback	Turn management	Semantic relations
Host vs. Guest				
Host			take, offer	substitution, contradiction
Guest	iconic, symbolic	give/non-accept	accept, yield	
Genres, host				
Political			take	contradiction
Cultural				substitution
Genres, guest				
Political		give/non-accept	yield	
Cultural	iconic		accept, complete	

the two subcorpora within the same speaker group (*Genres, host* and *Genres, guest* accordingly).

Semantic Relations. The last column of Table 3 refers to the relations of produced non-verbal expressions with the accompanied speech. Hosts seem to be more productive in the types of substitution and contradiction. Contradiction is a relation that appears only in the political corpus, and it includes cases where the host is ironic about what the politician says, thus, his/her non-verbal expressions contradict what is being said. Substitution, i.e. non-verbal expressions produced in the absence of speech, is mostly produced by hosts in the cultural corpus, and it is attributed to cases of silent feedback and backchanelling.

Emotions/ Attitudes. While expressing their emotions and attitudes, the hosts are more productive with regards to values such as annoyance, amusement, surprise and irony, all of which are responses to the guest's speech. Again, irony, annoyance as well as doubt are hosts' reactions observed mostly in political interviews. Anger, excitement and arrogance were found mostly on the guest's side. Guests are more prone to show their anger, worry and insecurity in the political corpus rather than in the cultural corpus. On the other hand, in the cultural corpus we attested a set of positive emotions such as excitement, affection and joy, which are very unlikely to occur in the political corpus.

Dialogue Acts. Finally, we also compared the dialogue acts that the speakers performed. In our study, we take into account the general-purpose functions as defined in the draft of ISO 24617-2 [5], with a reference to speech content only, since non-verbal aspects are taken care of the respective layers of the MUMIN scheme. Even so, though they are related to speech content and not to the non-verbal activity, dialogue act labels are good indicators of certain speaker profile traits. Acts such as request and suggestion are performed mostly by the hosts, while acts of confirmation, disconfirmation, justification and acceptance of requests are performed by the guests. It is notable that the quantity of acts expressing acceptance, confirmation and

Table 4. Statistically significant values in emotions/attitudes & dialogue acts features between speaker roles and between the two subcorpora

	Emotions/attitudes	Dialogue acts
Host vs. Guest		
Host	annoyance, amusement, surprise, irony	request, suggestion
Guest	anger, excitement, arrogance	confirmation, disconfirmation, justification, accept request
Genres, host		
Political	irony, doubt, annoyance	suggestion
Cultural	amusement	
Genres, guest		
Political	anger, worry, insecurity	disconfirmation
Cultural	excitement, affection, joy	confirmation

agreement is proportionally bigger in the cultural corpus, while acts of the opposite content (disconfirmation, disagreement) can be found more often in the political corpus.

4 Discussion

The interview genre in general is governed by certain rules and the situational and communicative context in which an interview is situated imposes specific restrictions in the conversational behavior of the participants. For example, a typical structure concerns the format of questioning and answering: a specific agenda of topics to be discussed exists, and the host is in charge of regulating the interaction. However, depending on the degree of formality, such behaviors may vary. In the case tackled in this paper, a mapping can be inferred between institutional interaction and political interviews, as well as less institutional, more casual interview types, which are represented by the cultural corpus. Moving from the strictly institutional (political) interviews towards the less formal, more casual ones, we observed that semiotic, feedback, turn management, feedback and emotional features are accommodated to fit the corresponding discourse type (e.g. narrative in the cultural dataset, argumentation in the political dataset) and support the speaker's intention and role within the specific communicative situation (e.g. to persuade, to give information, to disclose personal experiences, etc.).

Certain values are more predictable to occur than others as the communicative roles perform specific functions, which are annotated correspondingly: it is therefore normal for a host to employ non-verbal activity to e.g. offer the turn more often than the guest and the guest to accept the turn more often than the host respectively.

However, when such a comparison is made, we should be able to interpret not only the differences, but the similarities as well. Sometimes, similarities between speaker roles or genres might be against expectations, that is, a break in certain discursive or conversational rules. In those cases, further study of the context is needed. For example, the value of *disconfirmation* is expressed by both roles, a fact that indicates that in

particular parts of the interview the roles converge and the dimension of asymmetry, as imposed by the framework of interaction in an interview context, is absent.

The profiles of the speakers depend on the non-verbal expression that they produce as well as the semantics of those expressions. The semantics are defined by the preferences of the speakers, their communicative goals and obligations. Although the preferences sometimes converge to a large extent, there are subtle yet important differences. If the context imposes different goals and obligations, then the functions of the non-verbal behavior of the speakers are expected to serve and achieve those goals. For example, a major communicative goal in the political interviews is persuasion. Thus, strategies such as grabbing the turn from the co-locutor to support an argument, or expression of negative emotions when a politician feels misinterpreted by the host are very common and idiosyncratic in this interview type. As it is already mentioned, there are almost no differences in the quantity of the produced non-verbal expressions. Thus, although it is safe to generalize on quantity conclusions, this is not the case with regards to the quality features. There are subtle differences that are crucial to the interpretation of the behavior. In general, non-verbal features as expressed through turn management, feedback and emotions/ attitudes expression are closely linked to the context and therefore they explicitly influence the type of the semantic and pragmatic relationship developed between the verbal and non-verbal channels.

References

1. Ekman, P., Friesen, W.V.: The Repertoire of Nonverbal Behaviour: Categories, Origins, Usage and Coding. Semiotica 1, 49–98 (1969)
2. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press (2004)
3. McNeill, D.: Hand and Mind: What Gestures Reveal About Thought. University of Chicago Press, Chicago (1992)
4. Koutsombogera, M., Papageorgiou, H.: Multimodality Issues in Conversation Analysis of Greek TV Interviews. In: Esposito, A., Hussain, A., Marinaro, M., Martone, R. (eds.) COST Action 2102. LNCS (LNAI), vol. 5398, pp. 40–46. Springer, Heidelberg (2009)
5. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.R.: Towards an ISO Standard for Dialogue Act Annotation. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, MALTA (2010)
6. W3C Incubator Group: Emotion Markup Language – Working Draft (2009), http://www.w3.org/TR/2009/WD-emotionml-20091029/
7. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. Multimodal Corpora for Modeling Human Multimodal Behaviour. Journal on Language Resources and Evaluation 41(3-4), 273–287 (2007)
8. Heritage, J.: Conversation Analysis and Institutional Talk. In: Sanders, R., Fitch, K. (eds.) Handbook of Language and Social Interaction, pp. 103–146. Lawrence Erlbaum, New Jersey (2005)
9. Ilie, C.: Semi-institutional Discourse: The Case of Talk Shows. Journal of Pragmatics 33, 209–254 (2001)

The Neurophonetic Model of Speech Processing ACT: Structure, Knowledge Acquisition, and Function Modes

Bernd J. Kröger[1], Jim Kannampuzha[1], Cornelia Eckers[1], Stefan Heim[2,3,4], Emily Kaufmann[5], and Christiane Neuschaefer-Rube[1]

[1] Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital Aachen and RWTH Aachen University, Germany
{bkroeger,jkannampuzha,ceckers,cneuschaefer}@ukaachen.de
[2] Section Functional Brain Mapping, Department of Psychiatry, Psychotherapy, and Psychosomatics
[3] Section Neurological Cognition Research, Department of Neurology, University Hospital Aachen and RWTH Aachen University, Germany, and
[4] Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Germany
sheim@ukaachen.de
[5] Education and Rehabilitation of the Deaf and Hard of Hearing, University of Cologne, Germany
emily.kaufmann@uni-koeln.de

Abstract. Speech production and speech perception are important human capabilities comprising cognitive as well as sensorimotor functions. This paper summarizes our work developing a neurophonetic model for speech processing, called ACT, which was carried out over the last seven years. The function modes of the model are production, perception, and acquisition. The name of our model reflects the fact that vocal tract ACTions, which constitute motor plans of speech items, are the central units in this model. Specifically (i) the *structure* of the model, (ii) the acquired *knowledge*, and (iii) the correspondence between the model's structure and specific *brain regions* are discussed.

Keywords: neurophonetic model, speech processing, speech production, speech perception, speech acquisition, vocal tract actions, motor plan.

1 Introduction

Speech production and speech perception are important human capabilities comprising cognitive as well as sensorimotor functions. Realistic modeling of speech processing is an important part of understanding multimodal face-to-face interaction and thus of understanding important parts of social interactions. The neurophonetic model of speech processing presented in this paper comprises three function modes: speech production, speech perception, and speech acquisition. On the one hand, the model is based on a specific *neuroanatomical structure* for motor, sensory, and phonemic representations of speech [1, 2]. On the other hand, the model acquires *linguistic knowledge* as well as *speech skills* for a specific language. This knowledge and skills become integrated into the model based on synaptic weights (i.e. the degree

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 398–404, 2012.

of excitatory and/or inhibitory synaptic connections) between neurons of different neural maps [2]. Purely cognitive linguistic approaches are beyond the scope of this paper but a blueprint for integrating our model into a complete approach to speech processing including lexical representations is outlined in [3].

Basically, our neurophonetic model, in which *vocal tract ACTions* are assumed to constitute the basic units of motor plans (leading to the name ACT), is inspired by the organization of the previously only quantitative sensorimotor speech production model, i.e. the DIVA model [4, 5, 6]. Both approaches comprise sensorimotor feed-forward and feedback loops (Fig. 1). Starting from a phonemic representation, both approaches (DIVA and ACT) are capable of generating proper articulator movement patterns and subsequently proper acoustic speech signals. From the viewpoint of speech perception, it has been demonstrated that ACT is capable of modeling *categorical perception* [2, 7]; in this paper, we report on the model's ability to assign categorical perception to the topology of the *phonetic map*. A phonetic map is assumed to constitute the central supramodal neural map within a (language specific) speaking skills repository (i.e. vocal tract action repository [7]), and it associates motor, sensory, and phonemic states of speech items (Fig. 1).

2 The Structure of the Model and Its Function Modes of Production and Perception

The structure of our neurophonetic model is given in Fig. 1. *Speech production* starts with the phonemic representation of a speech item. This speech pattern (e.g. a word or a short utterance) is processed syllable by syllable. In the case of a *frequent syllable*, for which the motor plan has already been acquired, first the motor plan state is activated via the phonetic map, and subsequently the motor neuron activation pattern (level of the primary motor map) is generated for each vocal tract action occurring within the syllable. The subsequent neuromuscular processing leads to articulator movements and allows the generation of the acoustic speech signal via our articulatory-acoustic model [8]. The previously-acquired sensory state of this syllable is co-activated in parallel via the phonetic map (internal or trained state TS; Fig. 1). This state TS is matched with the state ES (external state ES; Fig. 1), resulting from the current production of that syllable. In the case of noticeable differences, an auditory and somatosensory error signal (Δau and Δss; Fig 1) is propagated via the phonetic map in order to alter the motor plan of that syllable for a new (corrected) production of that syllable. In the case of *infrequent syllables*, a motor plan is generated via the motor planning module by activating the plan of a phonetically and phonotactically similar syllable via the phonetic map [9].

Two control mechanisms are featured in our model. Firstly, *lower level compensatory corrections* occur in real time by correcting articulator position and velocity of articulators with respect to action goals as defined on the motor plan level (module of subcortical/cortical motor programming, execution and control; Fig. 1). On this level, compensation results from previously-acquired knowledge about possible vocal tract action realizations, especially if more than one articulator is involved (e.g. lower jaw and lips in the case of labial closing). This type of compensation has been exemplified

in bite block experiments [10] and in experiments introducing unexpected jaw perturbations [11].

Secondly, *sensorimotor adaptation* can be modeled in our approach by comparing internal (or trained) and external sensory states TS and ES (cortical level; Fig. 1). Basically, these states do not show noticeable differences (also called "error signals" Δau and Δss; Fig 1) after speech acquisition. But error signals can occur, for example if the lower-level perceptual processing system is modified artificially (e.g. by permanently shifting the second formant F2 via a specific real-time signal processing procedure [12]). The resulting adaptation effects have been explained in detail in the DIVA approach [4].

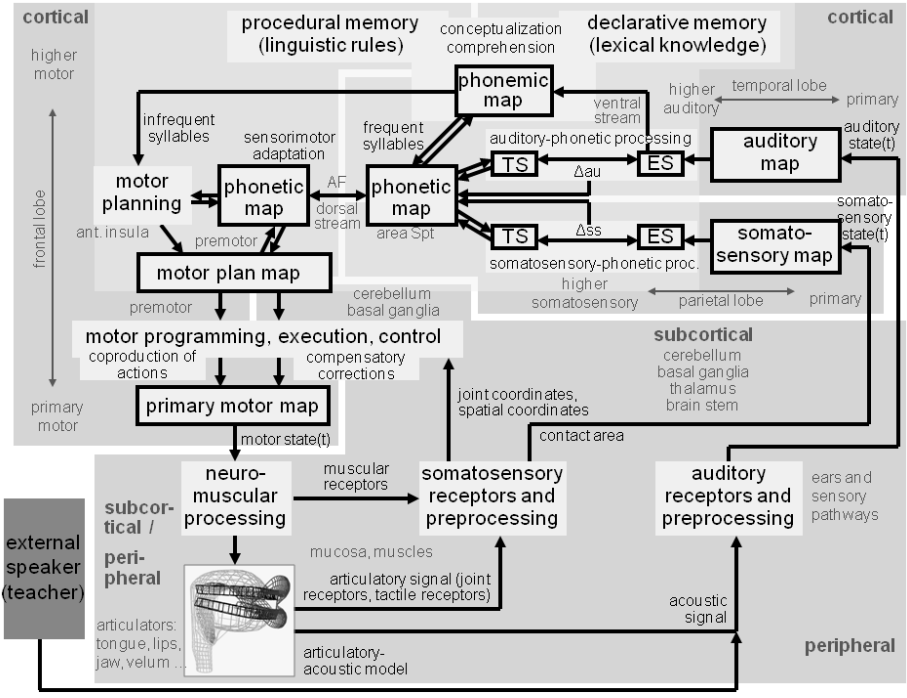


Fig. 1. Structure of the neurophonetic model ACT. Non-framed boxes indicate processing modules, framed boxes indicate neural maps. Single arrows indicate topology-preserving connections or streams (one-to-one mappings; parallel neural connections only); double arrows indicate complex neural mappings (all-to-all mappings; crossing neural connections). Dark grey: external speaker (mainly teacher or caretaker during speech acquisition); medium grey and light grey: the neural model; the medium grey region comprise modules and maps with temporal processing in short time intervals (12.5 msec in the current implementation of the model); the light grey region comprises modules and maps which process syllables as a whole unit (state maps within this region are part of short-term memory; mapping from state maps onto the phonetic map, as well as the phonetic map itself, are part of long-term memory). TS: trained or internal (sensory) states; ES: external (sensory) states; Δau: auditory error signal; Δss: somatosensory error signal; area Spt: area in the Silvian fissure at the parieto-temporal boundary [13]; AF: arcuate fasciculus.

Speech perception starts with an external acoustic signal. If phonemic identification is intended, the signal must be a realization of a frequent syllable. For this purpose the signal is preprocessed at peripheral and subcortical levels and loaded to the short-term memory as an external auditory state (ES; Fig 1). Then its neural activation pattern is passed to the trained state map (TS; Fig 1), firstly leading to the co-activation of a neuron region on the level of the phonetic map and secondly to the co-activation of a specific neuron on the level of the phonemic map; the first representing that syllable phonetically, the second phonologically. This neural pathway via the phonetic map, also referred to as the *dorsal stream* of speech perception [13], also co-activates a motor plan pattern for this frequent syllable. A second stream in speech perception, i.e. the *ventral stream*, directly links the auditory activation pattern with the phonological processing module [ibid.]. The dorsal stream is assumed to be important during speech acquisition, while the ventral stream is dominant later on during adult speech perception. The ventral stream is indicated in Fig. 1, but it has not yet been integrated into ACT.

3 Training Experiments for Acquiring Speech Knowledge and Speaking Skills

Our training experiments always comprise a *babbling phase* and an *imitation phase* [2] (see also DIVA model [4]). During babbling training, the model associates motor plan states with auditory states. On this basis, the model is capable of generating motor plan states during imitation training. We started with experiments on a phonotactically simple "*model language*" comprising V- and CV-syllables, with five vowels (V = /i/, /e/, /u/, /o/, /u/) and three consonants (C = /b/, /d/, /g/). All combinations of vowels and consonants occurred within the syllables with equal frequency [2]. We were able to show that on the level of the phonetic map, a strict phonetic ordering of realizations (exemplars) of these syllables occurred during babbling training (*phonetotopy* [14]). During imitation training, *phoneme regions* appeared on the level of the phonetic map [2, 7]. After these initial experiments we proceeded with a more complex model language which comprised V-, CV-, and CVV-syllables and which is based on a larger set of consonants (/b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /l/). The training once again resulted in a strict ordering of the phonetic map with respect to phonetic features (e.g. place and manner of consonant articulation) and phonotactic features such as syllable type (C, CV, or CCV) and types of consonants within the CC-cluster [9].

In the present experiments, we trained sets of the *200 most frequent syllables of Standard German* [15]. One of the most interesting results is that phonetic and phonotactic ordering in a real language is less strict than in the "model languages" we trained. This may be due to the fact that a real language is not constructed in a strictly regular way with respect to phonotactics, meaning not all CV or CCV combinations occur in a real language, or at least they do not occur with equal frequency. Furthermore, syllable exemplar regions are of different size in the case of a real language. Here, size is corresponding to the frequency of occurrence of a syllable in that particular language (Fig. 2). Thus in the case of this experiment up to 10 model neurons represent different realizations of a syllable.

Production and perception quality of the model was checked for the 50 most frequent syllables. Perception quality was quantified by the percentage of syllables, produced by a natural speaker and correctly identified by the model. This rate was 92% in the case that test items and training items were produced by the same speaker. The identification rate dropped to 84% if syllables were produced by a different male speaker. In the case of the same speaker, test and training items were different, i.e. chosen from different subsets of syllable realizations. Production quality was quantified by the percentage of syllables, correctly identified by natural subjects (listeners). The 50 most frequent syllables were produced by the model. Perception tests were performed by 5 persons between 25 and 28 years old with no known speech perception deficits. The mean rate of correct identifications was 96%.

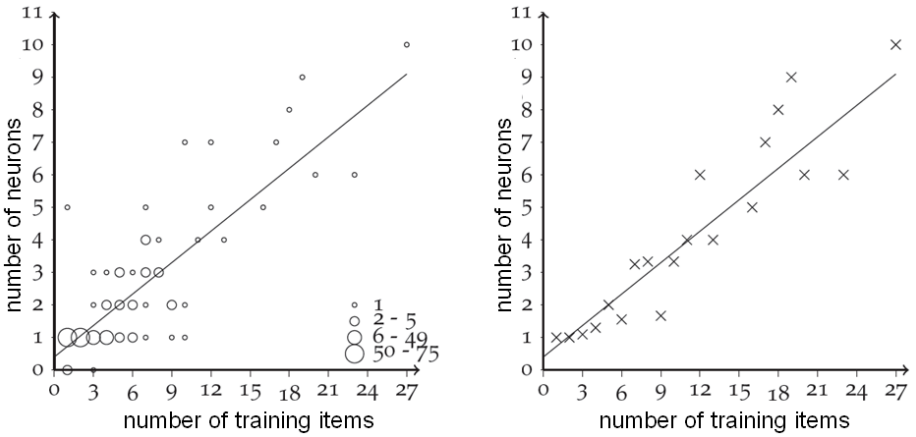


Fig. 2. Number of neurons representing a certain syllable as functions of number of training items which are used for training of that certain syllable. Left side: the size of the circle represents the number of different syllables exhibiting the same number of training items and leading to the same number of neurons representing that specific syllable; Right side: arithmetic mean of neurons representing all syllables with an identical number of training items. The training was performed for 200 most frequent syllables of Standard German uttered in sentence context uttered by a 33 year old male speaker without any known speaking or hearing abnormalities [15].

4 Neuroanatomical Correlates

In order to specify the neuroanatomical correlated neural maps, mappings and pathways needed to be differentiated. *Neural maps* comprise all state maps (i.e. phonemic, auditory, somatosensory, motor plan, primary motor maps; Fig. 1) as well as the self-organizing map (i.e. the phonetic map; see Fig. 1). *Neural mappings* occur between the self-organizing map and all state maps (double arrows in Fig. 1). *Neural pathways* occur between maps or between a map and a processing module (single arrows in Fig. 1). In contrast to neural mappings neural pathways are not capable of *generating new* neural activation pattern, but rather *forwards* an already existing

neural activation pattern from one map to another. Thus, neural streams are represented as bundles of *parallel neural fibers* and are capable of connecting non-adjacent brain regions (e.g. arcuate fasciculus [16]), while neural mappings are realized by complex all-to-all cross-connection networks. Mappings mainly connect neural maps which are in close range of each other.

The short-term memory state maps as defined in our model (state maps within the light grey area in Fig. 1) comprise a temporal storage over the time interval of at least one syllable [2, 3]. These higher-level state maps are assumed to be located near the brain regions of the sensory error maps postulated in [4, 5]. In contrast, the lower-level state maps (state maps within the medium grey area in Fig. 1), process the stream of motor data for articulation and the stream of sensory data, already preprocessed by peripheral modules. These state maps are located in primary sensory and motor areas (Fig. 2).

A further comparison of the structure of our model to the structure of the speech processing model proposed by Hickok and Poeppel [13] leads to the conclusion that the supramodal phonetic map cannot be located in *one* particular brain region. Moreover, it is assumed that the phonetic map is copied from the area in the Silvian fissure at the parieto-temporal boundary (labeled as area Spt [13]) onto the premotor area by a neural stream (i.e. by the arcuate fasciculus) in order to allow close-range neural mappings between the phonetic map and the motor plan state map, on the one hand, and between the phonetic map and the phonemic, auditory, and somatosensory short-term memory state maps on the other hand (Fig. 1). Thus the phonetic map could be interpreted as a mirror neuron system at the phonetic level in contrast to the well-known semantic level mirror neurons [17].

5 Discussion and Further Work

Although both quantitative sensorimotor models of speech production and speech processing in principle are compatible, what makes our neurophonetic model ACT [2, 7, 9, 14, 15] different from the DIVA model [4, 5, 6] is that it allows an alternative view on the neurophonetics of speech production, perception, and acquisition. While the DIVA model mainly focuses on exemplifying sensorimotor adaptation [4, 5], our model focuses on exemplifying the development of a vocal tract action repository (i.e. phonetic map) as the central repository for sensorimotor speech knowledge and speaking skills on the basis of principles of neural self-organization [2, 14]. Brain imaging experiments are planned in order to verify or falsify our hypotheses, especially on the mirror-neuron-character of the phonetic map. Further modeling experiments are planned in order to gain more insight into the development of phonetic knowledge and phonological structure of a specific language within a complete neurolinguistic approach. This approach would also include the development of the mental lexicon along with the development of the vocal tract action repository [3].

Acknowledgments. This work was supported in part by German Research Council (DFG) grants Kr 1439/13-1 and Kr 1439/15-1 and in part by COST-action 2102.

References

1. Kröger, B.J., Kopp, S., Lowit, A.: A model for production, perception, and acquisition of actions in face-to-face communication. Cognitive Processing 11, 187–205 (2010)
2. Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C.: Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793–809 (2009)
3. Kröger, B.J., Birkholz, P., Neuschaefer-Rube, C.: Towards an articulation-based developmental robotics approach for word processing in face-to-face communication. PALADYN Journal of Behavioral Robotics (in press)
4. Guenther, F.H.: Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39, 350–365 (2006)
5. Guenther, F.H., Ghosh, S.S., Tourville, J.A.: Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language 96, 280–301 (2006)
6. Guenther, F.H., Vladusich, T.: A neural theory of speech acquisition and production. Journal of Neurolinguistics (in press)
7. Kröger, B.J., Birkholz, P., Kannampuzha, J., Neuschaefer-Rube, C.: Categorical Perception of Consonants and Vowels: Evidence from a Neurophonetic Model of Speech Production and Perception. In: Esposito, A., Esposito, A.M., Martone, R., Müller, V.C., Scarpetta, G. (eds.) COST 2102 Int. Training School 2010. LNCS, vol. 6456, pp. 354–361. Springer, Heidelberg (2011)
8. Kröger, B.J., Birkholz, P.: A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) Verbal and Nonverbal Commun. Behaviours. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)
9. Kröger, B.J., Miller, N., Lowit, A.: Defective neural motor speech mappings as a source for apraxia of speech: Evidence from a quantitative neural model of speech processing. In: Lowit, A., Kent, R. (eds.) Assessment of Motor Speech Disorders, pp. 325–346. Plural Publishing, San Diego (2011)
10. Fowler, C.A., Turvey, M.T.: Immediate compensation in bite-block speech. Phonetica 37, 306–326
11. Kelso, J.S., Tuller, B., Vatikiotis-Bateson, E., Fowler, C.A.: Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for cooperative structures. Journal of Experimental Psychology: Human Perception and Performance 10, 812–832 (1984)
12. Houde, J.F., Jordan, M.I.: Sensorimotor adaptation of speech I : Compensation and adaptation. Journal of Speech, Language, and Hearing Research 45, 295–310 (2002)
13. Hickok, G., Poeppel, D.: Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4, 131–138 (2007)
14. Kröger, B.J., Kannampuzha, J., Lowit, A., Neuschaefer-Rube, C.: Phonetotopy within a neurocomputational model of speech production and speech acquisition. In: Fuchs, S., Loevenbruck, H., Pape, D., Perrier, P. (eds.) Some Aspects of Speech and the Brain, pp. 59–90. Peter Lang, Berlin (2009)
15. Kröger, B.J., Birkholz, P., Kannampuzha, J., Kaufmann, E., Neuschaefer-Rube, C.: Towards the Acquisition of a Sensorimotor Vocal Tract Action Repository within a Neural Model of Speech Processing. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 287–293. Springer, Heidelberg (2011)
16. Bernal, B., Ardila, A.: The role of the arcuate fasciculus in conduction aphasia. Brain 132, 2309–2316 (2009)
17. Rizzolatti, G., Craighero, L.: The mirror-neuron system. Annual Review of Neuroscience 27, 169–192 (2004)

Coding Hand Gestures: A Reliable Taxonomy and a Multi-media Support

Fridanna Maricchiolo[1], Augusto Gnisci[2], and Marino Bonaiuto[3]

[1] Dipartimento di Studi dei Processi Formativi Culturali e Interculturali nella Società Contemporanea, Università degli Studi Roma Tre, via Milazzo 11b 00185, Rome, Italy
[2] Dipartimento di Psicologia, Seconda Università degli Studi di Napoli
[3] Dipartimento di Psicologia dei Processi di Sviluppo e Socializzazione, Sapienza Università di Roma, Italy
fmaricchiolo@uniroma3.it, augusto.gnisci@unina2.it, marino.bonaiuto@uniroma1.it

Abstract. A taxonomy of hand gestures and a digital tool (CodGest) are proposed in order to describe different types of gesture used by speaker during speech in different social contexts. It is an exhaustive and mutually exclusive categories system to be shared within the scientific community to study multimodal signals and their contribute to the interaction. Classical taxonomies from gesture literature were integrated within a comprehensive taxonomy, which was tested in five different social contexts and its reliability was measured across them through inter-observer agreement indexes. A multi-media tool was realized as digital support for coding gestures in observational research.

Keywords: coding gesture, multi-media tool, reliability, observational research.

1 Introduction

Research on hand gesture falls within the behavioural analysis of interactions. Behavioural studies often involve observational methods, i.e., the adoption and/or adaptation of reliable coding systems shared by the scientific community [1]. Nevertheless, within gesture literature, a number of gesture types and different classifications of them have been offered, e.g., [2-5]; Wundt, 1921/1973, in [6], along different dimensions/criteria. The focus of this article is to propose a reliable coding system based on a taxonomy of hand gestures used for observation of social interaction within different social contexts and situations. To this end, a multi-media manual as coding support is proposed.

Up to now, there is no a universally shared category system for hand gestures. Many authors differently described the variety in which gestures occur and are used by the speaker. These differences are also evident in the criterion/a at the basis of each taxonomy, that probably depend on the scientific discipline their authors refer to and on the aims they pursue.

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 405–416, 2012.

Ekman and Friesen [3], using three taxonomic criteria (usage, origin, and coding), distinguished five main categories of gestures: Illustrators, conveying semantic content; Emblems, conventional and cultural signs; Regulator signals, controlling conversational flow; Emotional displays, expressing emotional states; and Adaptors, contact and manipulation hand movements. Kendon [4] referring to Gesticulation (the gesture that is incomplete without speech accompaniment) used a continuum criterion: from Spontaneous Gesticulation, Language-slotted, Pantomime, Emblems, to Signs. McNeill [5] distinguishes gestures belonging to ideation process (propositional gestures, representing linguistic referents: iconics, metaphorics, and deictics) and gestures characterizing discursive activity (non-propositional gestures: cohesive and beats). Bavelas and coll. [2] distinguish: Topic gestures, i.e., referential or semantic gestures (similar to illustrators) and Interactive gestures, having an intrinsically interpersonal character (often represented by pointing to the addressee like deictics). Krauss and coll. [7] distinguish hand movements on a continuum of lexicalization (level of resemblance or closeness with the words): from Symbolic gestures, with a higher degree of lexicalization, to Conversational gestures, distinguished into Lexical gestures (connected with the semantic content of the speech) and Motor movements (coordinated to the prosody of the speech) with a medium degree of lexicalization, up to Adapters (communicatively meaningless and not related to speech) with a lower lexicalization degree. Poggi [8] distinguishes communicative gestures according to various criteria. One of these is the cognitive construction: codified *vs*. creative gestures. The first are meaningful signs steadily represented in memory (e.g., emblems). Creative gestures are performed to represent or evocate some referent (e.g., iconic, but also deictic gestures). According to the author, creative gestures are motivated, while the codified ones, though having an iconic dimension, are often arbitrary. Not all these classifications are exhaustive. Moreover, they often adopt different taxonomic criteria: in some cases there is not a single criterion as there are more than one [3, 8]; in some cases the categories are not structured in specific sub-categories [2, 8] as they are simplifications of categories described by previous authors [9]. Some contributions only develop gesture annotation schemes [10] that maintain temporal structure and location information for capturing the original gestures and replicating them on an animated character. Allwood and colleagues [11] developed the MIMUN, a multimodal annotation scheme dedicated to the study of gestures (hand gestures and facial signals), with particular regard to their function about feedback, turn management and sequencing; gestures are also described through features of their shape and dynamics. Classifying a behavioural phenomenon, such as hand gesture, in a system of mutually exclusive and exhaustive categories is desirable as well as using coding systems shared by the literature, which proved to be reliable [12]. Achieving the usability of a taxonomy for making it available and widespread to the scientific community needs both conceptual and operational definition of each specific category as well as its concrete and exact description. Moreover, to establish a classification system it is necessary an empirical checking through behavioural analysis of interactions, where gestures are naturally occurring, as well as a coding of the different specific categories across several real interactive contexts. Then, in order to check if the proposed categories of gesture are recognized in a reliable way by the coders it is necessary to develop inter-observer agreement indexes such as Coehn's *K* [13] or Krippendorff's α [14]. Furthermore,

since nowadays the researcher work is mainly done through computer, a digital multi-media tool would be useful to make the behavioural analysis and coding job easier and straightforward. In spite of these arguments, studies on gesture classification do not always include, within a unique contribution, these minimal requirements, i.e.: empirical behavioural observation within different social situations and statistical analysis for reliability evaluation, as well as multi-media supports.

As a consequence, the general aim of this contribution is an hand gesture taxonomy and a multi-media tool for coding in observational research. Specific aims are:

1) definition and illustration of the proposed taxonomy and of gesture categories;
2) establishment of criteria for assessing if trained-to-the-gesture-taxonomy observers code in a reliable way many interactions from different contexts;
3) description of a digital manual for coding gesture.

Each of the next three paragraphs addresses one of the three aims.

2 A Hand Gesture Taxonomy for Observational Research

The taxonomy presented combines classical gestures categories (mainly, [3, 5]; but also, in some way, [2, 7]). Many authors (e.g., amongst others, [15, 5]) stressed the importance of gestural-verbal co-occurrence in both theoretical and methodological terms. Referring to this position, the fundamental criterion of the proposed taxonomy is that gestures can, or cannot, be linked to the speech. The term "link" here is intended not in terms of gesture-speech "co-occurrence", but with reference either to the "content" of the speech or to the "structure" of the speech. The taxonomy is organized along three hierarchical levels (Fig. 1): macro-, specific, and sub-categories. As reported in Fig. 1, according to the taxonomy, gestures are divided in Speech Linked Gestures (SLG) and Speech Non-linked Gestures (SNG). The SLG macro-category includes "cohesive" [5, 16], "rhythmic" [5, 16], and "ideational" gestures (specific categories). Cohesive category is distinguished in specific gestures (sub-categories), named according to the specific movement shape realized in the air by the hand(s) (e.g., "weaver", "whirlpool", "nipper", etc.). Ideational category includes "emblems" [3, 15] and "illustrators" [3]; illustrators, in turn, includes "iconic", "metaphoric", and "deictic" [16] gestures. SNG are distinguished in "self-adaptor gestures" [3, 17] and "hetero-adaptor gestures" [3, 18], which can be "person-addressed adaptors" and "object-addressed adaptors".

The conceptual definitions (CD) of a gesture category - explaining the theoretical notion (i.e., its correspondence with previously given definitions in literature) – and the operative ones (OD) - explaining observable elements useful for gesture recognition (i.e., how to observe and recognize the gesture) - are synthetically reported in Appendix. Since it is impossible to classify the gestures without considering the context and the function of the gesture [8, 19], such indications are given in the ODs.

Keeping the names already used for a very long time in the literature can be the best solution to share and to connect the present work within the scientific community.

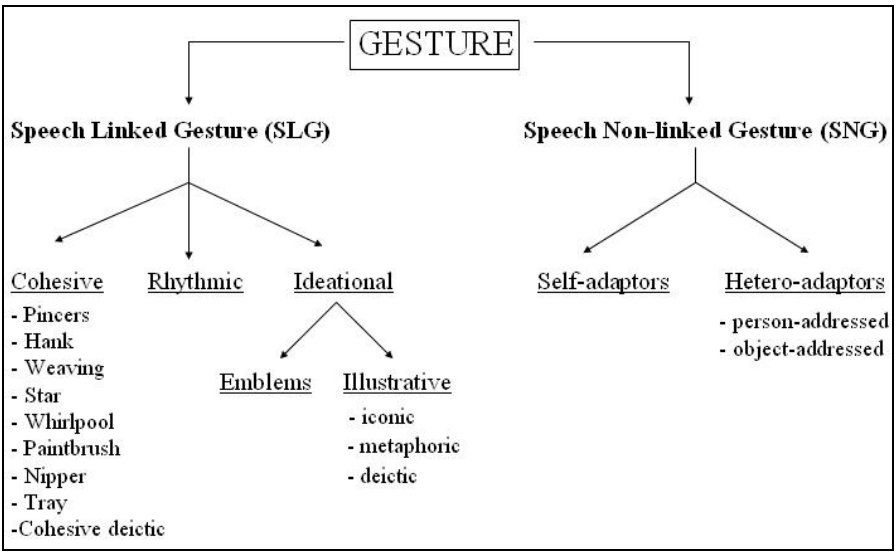


Fig. 1. Coding tree of the proposed taxonomy

3 Evaluating Reliability in Different Interactive Contexts

Hand gestures performed by persons interacting within different social contexts were observed and coded according to the above gesture categories system by trained coders. The sample of video-recorded social contexts has been selected amongst the archives of videotapes collected in many years at the Social Psychology Laboratory (Department of Psychology of Development and Socialization Processes, Sapienza University of Rome).

The following natural contexts (broadcasted by Italian TV) were considered:

1) television political interviews: two Italian political leaders of opposite alliance (Silvio Berlusconi and Francesco Rutelli, *N*=2), interviewed separately during the campaign for the 2001 Italian political elections;
2) courtroom examinations: a witness called by accuse (*N*=1) testifying against the man suspected to having killed her adoptive son.

The following laboratory contexts were considered:

3) simulation of small group discussion: four subjects (*N*=4) simulated to be members of an advisers group having to discuss two "business cases" to find one unanimous written solution for each case [20];
4) simulations of dyadic discussion: two subjects (*N*=2) had the same task of the group simulation (see above);
5) simulations of examination: five subjects (*N*=5) are singularly interrogated by a confederate within a simulated context where they have the task to answer in one part by telling the truth and in another part by lying [21].

All the subjects in the simulations were blind to the objectives of this research. In order to have comparable data across the settings, only 20 minutes from each context

have been selected for a total of 100 minutes. Interactions were transcribed by an adapted conversational analysis system [22] integrated with alternate lines in which gestures were signed. The observer proceeded to annotate gestures directly in the transcripts of the speech of the interactions while watching the videotapes. Speech-gesture synchrony observation leads to a degree of accuracy that permits assessment of how meaningful gestural movements co-occur with speech, syllable by syllable [19]. The segmentation (beginning and end) of each gesture was marked with square brackets as overlap on the verbal transcripts. In this way it was possible to segment gestures anchoring on the speech and to fix all the co-occurrences between spoken language and gestures. Hand gestures produced by the speakers during the videotaped interaction were coded according to the taxonomy categories. A coder "blind" to the research aims had preventively been trained by an expert coder (the first author of this chapter) to the use of the coding system and to the identification of each gesture category. Under her supervision, the observer examined various videotapes (working on a different sample of transcripts) and she was trained in recognizing all the different categories of the taxonomy. Only when the coder showed herself to be capable of coding gestures according to the coding system, she began to codify the whole selected material for the empirical test of the hand gesture taxonomy. Having finished the coding, the observer checked her own codings by observing all the video recorded material again. Any remaining ambiguous case was resolved in a discussion with her coding supervisor.

The frequency and percentage of each gesture category in each context is shown in Table 1.

Table 1. Amount and percentage (for context total) of gestures observed within five different contexts for each gesture specific category of the proposed taxonomy

Gestures	**Political**		**Courtroom**		**Group**		**Dyads**		**Simulated**		*Total*	
	f	**%**	**f**	**%**	**f**	**%**	**f**	**%**	**examination**		**f**	**%**
Cohesive	140	20.6	54	10.5	205	20	170	24	235	22.9	804	20.3
Rhythmical	70	10.3	93	18.1	138	13.4	30	4.2	59	5.8	390	9.9
Emblem	19	2.8	106	20.6	83	8.1	48	6.8	76	7.4	332	8.4
Iconic	2	0.3	5	1	2	0.2	3	0.4	58	5.7	70	1.8
Metaphoric	108	15.9	90	17.5	207	20.2	61	8.6	141	13.8	607	15.3
Deictic	266	39.1	102	19.8	159	15.5	113	15.9	137	13.4	777	19.6
Illustrator	376	55.3	197	38.2	368	35.8	177	25	336	32.8	1454	36.8
Self-adaptor	16	23.5	38	7.4	94	9.1	215	30.3	268	26.2	631	15.9
Object-ad.	59	8.7	29	5.6	137	13.3	68	9.6	50	4.9	343	8.7
Person-ad.	0	0	0	0	2	0.2	2	0.3	0	0	4	0.1
Tot. adaptor	75	11	65	12.6	233	22.7	284	40	318	31.1	975	24.6
Total	680	100	515	100	1027	100	709	100	1024	100	3955	100

Results demonstrate that almost all the gesture categories of the taxonomy are present in each different social context. It is possible to note a sort of differentiation in the use of hand gestures within the different social situations: some gesture categories occur more often in particular social contexts than in others. But some categories (i.e., illustrators) are used more often than all other categories across all the interaction contexts.

The results also demonstrate that it is possible to recognize and code all hand gestures observed during whatever kind of interactions using the gesture categories provided by the taxonomy: the category system is exhaustive.

To reach more efficient outcomes especially in terms of mutual exclusiveness, it has been necessary to evaluate the reliability of the coding system used for these observations and coding. For this purpose inter-rater agreement indexes on the general coding system for each context were calculated. Systematic observation and codification of hand gestures were carried out also by another independent observer, separately trained to use the coding system. The two observers (O_1 and O_2) separately and independently carried out the coding of the whole video-recorded sampled material, according to the above described procedure. Statistical data analyses were carried out to evaluate taxonomy reliability, measuring the concordance between two independent observers across different contexts. The percentage of agreement on gesture segmentation, the inter-rater agreement on the whole coding system of gestures in general and separated for each context were calculated by means of Cohen's *K* [13] (using the software ComKappa [23]). According to [1], Cohen's *K* >.75 is set as the threshold to test coding reliability. Given the fact that the length of each interaction is the same (20 minutes), simply frequencies of each coding category were used as the unit of analysis.

The percentage of agreement between O_1 and O_2 on recognition of a gesture (reliability on unityzing) is 91.6%, thus, excellent [24]. The total agreement on the whole coding system is $K = .82$, which can be considered good (K>.75). The *K* indexes were calculated separately on each specific gesture category and on the whole gesture category system for each observed context. The minimum value of *K* is.75 for each of the eight specific gesture categories and for the whole system in each of the five contexts considered. Gestures more reliably identified by coders are Iconics and Self-adaptors (K=.94) while gestures less reliably identified are Rhythmics and Metaphorics (K=.75). Similarly, for the context "Simulated interrogatory" *K* index turns out to be the lowest: nevertheless it is still acceptable even according within a conservative approach such as the one by Bakeman and Gottman's standard. [1].

Results as a whole demonstrate that the categories of the proposed taxonomy can be recognized, discriminated and identified in reliable way in different interaction contexts.

Since the intercultural debate on gestures focuses on culture-specificity of gesture [25], we have begun a first step toward an inter-cultural validation of the taxonomy. The presented taxonomy has been tested in dyadic interactions (two young women talking about life and work problems) among members of another, very different culture from the Italian one (under many respect, such as economic, religious, social, ethnic, education conditions): they were women in Burkina Faso. Moreover these persons interacted using two different languages in two different moments [26, 27]. Further publications are in progress to report full details for that research. Observation of gestures of Burkinabe peoples have been compared to Italian ones (in the same type of conversations), finding that the same gesture categories (but in different amount) of the taxonomy have been observed also in non-Italian people. Such a test demonstrates that this system is based on abstract categories and not on culture-specific functions. In different cultures, however, gesturer varies for the amount of frequency (in each category), for the entity of movement, for the space use, and for the types and meaning of emblems [25].

4 Multimedia Tool for Coding Hand Gestures - CodGest

A multi-media manual (presently only in Italian language), called CodGest, has been realized as a tool to: describe the gesture taxonomy; support observational coding; make it shared by other researchers. Upon request, CodGest manual is available from the Authors. This instrument offers audiovisual support to gesture study and, in particular, for learning and using the gesture category system, in order to make such a taxonomy more easy to be consulted and used during observation. The tool is developed on digital interactive multimedia support: normal text is thus integrated by important information in form of images as well as audio-video, reproducing speakers performing hand gestures. Texts, images and videos were assembled through *Macromedia Flash MX*. CodGest is composed as follows:

a) a home page and a brief theoretical introduction, with hypertext, summarizing salient points and basic principles on which the taxonomy is based, with literature bibliographical references;
b) a brief paragraph for each gesture category, including both a conceptual and an operational definition, as well as a verbal description of shape and movement(s) performed by hand(s) (see Appendix);
c) three examples ("ideal", "prototypical", "problematic") of each gesture category, in video images taken one from *ad hoc* videos ("ideal" examples built with actors for this end) and two from "field" samples (see paragraph 3: both "prototypical", i.e., clear, and "problematic", i.e., dubious, examples);
d) an example of gesture, for each category, through a three- or four-picture sequence, realized *ad hoc* in the laboratory and aimed at showing prototypical shape and movement of gesture;
e) coding notes to facilitate and resolve possible problems in assigning a code to each gesture.

CodGest referring to the verified category system was developed in digital format but in compliance with traditional methodological criteria for coding manual realisation [1]: a phenomenon description articulated in conceptual and operative definitions of different categories, with ideal, real, typical and problematic examples of them. The advantage of this is that the digital support in multi-media interface permits the addition of important information, such as photographic and audio-visual examples, to the text: these are fundamental for completely and appropriately understanding any coding system, as well as for sharing it. Some authors, in fact, maintain that in observational research it is desirable that the scientific community has at its disposal shared tools, allowing the researcher to compare own data and outcomes with that of others [12].

5 Conclusion

The coding system of hand gestures proposed in this study integrates and synthesizes main existing gesture classifications (in particular, [2, 3, 5, 16]), with the aim of individuating a general taxonomy useful to recognize and code hand gestures within a range of social situations. The observation and coding of hand gestures carried out within different contexts and the consequent individuation of an amount of occurrence

for each gesture category in each context demonstrates the usefulness of the proposed taxonomy to study hand gestures in different dyadic or group interactions. The inter-observers agreement indexes calculated in order to measure the reliability of the way in which the category of coding system can be recognized, in general, turn out to be very satisfactory. These results allow the use of this taxonomy as a good tool for coding hand gestures in the study of interaction in various social contexts. Different authors maintain that in observational research it would be desirable that the scientific community shares tools in order to permit the researchers to compare one's own data as well as one's own outcomes with the other's ones [12]; and this is particularly true for social interaction researches which heavily rely upon observational techniques. In some fields of bodily communication such a tool partly exists (e.g., facial expression and recognition of emotion), while in other fields they are mostly lacking. The multimedia support, CodGest, for hand gesture coding represents a first step along this direction within the field of hand gestures. Such an instrument offers different possibilities under the forms of images, videos and texts to know and recognize specific categories of gesture. An important aspect of this study is that the same system of gesture categories had tested in five different social contexts, contrary to previous studies in the field which almost exclusively focused on a single category or context each time [2, 28-31]. The high scores of agreement indexes calculated in all the contexts confirm not only that the categories can be recognized in a reliable way, but also its appropriateness and effectiveness as a tool for observation and study of gestures in various social situations, implying some stability in the "kind" of gestures used across different settings of social interaction. This result gives a sort of generalization to the outcomes here obtained. However, it requires to be submitted to a more thorough verification via a generalizability analysis [32] in order to estimate if and how much this category system discriminates between subjects or between contexts (between variance) and between gesture categories, rather than between observers (within variance).

Another aspect of generalization and validation is cross-cultural comparison, which has already been carried out by our research team with good outcomes [26, 27]. A further development consists in analyses aimed at checking each category function through their co-occurrence with specific verbal phenomena and devices [33].

A planned improvement to develop the digital manual is an exercise section for the user, aimed at training him/her to use the taxonomy and, therefore, to calculate the reliability of his/her measure with reference to a coding standard protocol. In this way it will be possible to compare the reliability of observers trained through digital manual support with the reliability of observers trained through a traditional method (i.e., via only textual manual and agreement with researchers). Further taxonomical developments should certainly address more detailed issues, such as, for example, the movement components or parameters characterizing each gesture category or sub-category: therefore, more fine grained analyses could be usefully enclosed within the presented hand gesture coding system [34, 6]. Similarly the presented specific categories could be enclosed within a higher order hierarchical level of coding, i.e., more general concepts such as that of "family of gesture" [6]. This means that the presented categories could lie at an intermediate level of abstraction, having above the "family of gesture" coding level, and below the level of the single movements, parameters or phases composing each gesture's single enactment.

References

1. Bakeman, R., Gottman, J.M.: Observing interaction. An introduction to sequential analysis, II edn. Cambridge University Press, New York (1997)
2. Bavelas, J.B., Chovil, N., Lawrie, D.A., Wade, A.: Interactive gestures. Discourse Processes 15, 469–489 (1992)
3. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior. Semiotica 1, 49–98 (1969)
4. Kendon, A.: Gesture and speech: How they interact. In: Wiemann, J.M., Harrison, R.P. (eds.) Nonverbal Interaction, pp. 13–45. Sage Publications, Beverly Hills (1983)
5. McNeill, D.: Hand and Mind. The University of Chicago Press, Chicago (1992)
6. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge (2004)
7. Krauss, R.M., Chen, Y., Chawla, P.: Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? Adv. Exp. Soc. Psychol. 28, 389–450 (1996)
8. Poggi, I.: Iconicity in different types of gestures. Gesture 8, 45–61 (2008)
9. McNeill, D.: So you think gesture are nonverbal? Psychol. Rev. 92, 350–371 (1985)
10. Kipp, M., Neff, M., Albrecht, I.: An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. Lang. Resour. Eval. 41(3-4), 325–339 (2007)
11. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Lang. Resour. Eval. 41(3-4), 273–287 (2007)
12. Bakeman, R., Gnisci, A.: Sequential observational methods. In: Eid, M., Dieneer, E. (eds.) Handbook of Multimethod Measurement in Psychology, pp. 127–140. American Psychological Association, Washington (2006)
13. Cohen, J.A.: Coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46 (1960)
14. Krippendorff, K.: Estimating the reliability, systematic error, and random error of interval data. Educ. Psychol. Meas. 30, 61–70 (1970)
15. Kendon, A.: Gestures as illocutionary and discourse structure markers in Southern Italian conversation. J. Pragmatics 23, 247–279 (1995)
16. McNeill, D., Levy, E.T.: Cohesion and gesture. Discourse Processes 16, 363–386 (1993)
17. Rosenfeld, H.M.: Instrumental affiliative functions of facial and gestural expressions. J. Pers. Soc. Psychol. 4, 65–72 (1966)
18. Edelmann, R.J., Hampson, S.: Embarrassment in dyadic interaction. Soc. Behav. Personal. 9, 171–178 (1981)
19. Duncan, S.: Coding manual (2004), Technical Report availale from `http://www.mcneilllab.uchicago.edu`
20. Maricchiolo, F., Livi, S., Bonaiuto, M., Gnisci, A.: Hand gestures and perceived influence in small group interaction. Span. J. Psychol. 14, 755–764 (2011)
21. Caso, L., Maricchiolo, F., Bonaiuto, M., Vrij, A., Mann, S.: The impact of deception and suspicion on different hand movements. J. Nonverbal Behav. 30, 1–19 (2006)
22. Jefferson, G.: On the interactional unpackaging of a "Gloss". Lang. Soc. 14, 435–466 (1985)
23. Robinson, B.F., Bakeman, R.: ComKappa: A Windows 95 Program for Calculating Kappa and Related Statistics. Behav. Res. Meth. Ins. C. 30, 731–732 (1998)

24. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. 76, 378–382 (1971)
25. Kita, S.: Cross-cultural variation of speech-accompanying gesture: A review. Lang. Cognitive Proc. 24, 145–167 (2009)
26. Bonaiuto, M., Gnisci, A., Maricchiolo, F.: Struttura e funzioni dei gesti delle mani durante la conversazione: nodi concettuali e possibili sviluppi futuri. Presentation to "Gesture in the Mediterranean: Recent Research in Southern Europe", Procida, October 21-23 (2005)
27. Bonaiuto, M., Maricchiolo, F., Orlacchio, T.: Cultura e gestualità delle mani durante la conversazione: un confronto tra donne native dell'Italia e del Burkina Faso. Presentation to Workshop: "Intersoggettività. Identità e Cultura", Urbino, October 14-15 (2005)
28. Beattie, G., Shovelton, H.: Iconic hand gestures and predistability of words in context in spontaneous speech. Brit. J. Psychol. 91, 473–492 (2000)
29. Beattie, G., Shovelton, H.: An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. Brit J. Psychol. 93, 179–192 (2002)
30. Contento, S., Stame, S.: Déixis verbale et non verbale dans la construction de l'espace interpersonnel. Dialogue Analisis 5, 427–433 (1997)
31. Feyereisen, P., Havard, I.: Mental imagery and production of hand gestures while speaking in younger and older adults. J. Nonverbal Behav. 23, 153–171 (1999)
32. Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N.: The dependability of behavioural measurement: Theory of generazability for scores and profiles. Wiley, New York (1972)
33. Maricchiolo, F., Bonaiuto, M., Gnisci, A.: Hand gestures in speech: studies on their roles in social interaction. In: Mondada, L. (ed.) Proceedings of the 2nd ISGS Conference, Interacting Bodies, Lyon-France, June 15-18. Ecole Normale Supérieure Lettres et Sciences humaines, Lyon (2007)
34. Calbris, G.: Elements of Meaning in Gesture. John Benjamins Publishing, Amsterdam (2011)
35. Wiener, M., Devoe, S., Rubinow, S., Geller, J.: Nonverbal behavior and nonverbal communication. Psychol. Rev. 79, 185–214 (1972)
36. Barakat, R.: Arabic gestures. J. Pop. Cult. 6, 749–792 (1973)
37. De Jorio, A.: La mimica degli antichi investigata nel gestire napoletano (Ancient's mimicry investigated through gesturing Neapolitan). Fibreno, Napoli (1832)
38. Ricci Bitti, P.E., Poggi, I.A.: Symbolic nonverbal behavior: Talking through gestures. In: Feldman, R.S., Rimé, B. (eds.) Fundamentals of Nonverbal Behavior, pp. 433–457. Cambridge University Press, New York (1991)
39. Morris, D.: Manwatching. A field guide to human behavior. Andromeda Oxford Limited and Jonathan cape Limited, London (1977)
40. Haviland, J.B.: Pointing, gesture spaces, and mental maps.. In: McNeill, D. (ed.) Language and Gesture, pp. 13–46. Cambridge University Press, Cambridge (2002)
41. Rosenfeld, H.M.: Instrumental affiliative functions of facial and gestural expressions. J. Personal. Soc. Psycho. 4, 65–72 (1966)
42. Freedman, N., Hoffman, S.P.: Kinetic behavior in altered clinical states: approach to objective analysis of motor behavior during clinical interviews. Percept. Motor Skill. 24, 527–539 (1967)
43. Kimura, D.: Manual activity during speaking. Neuropsychologia 11, 45–55 (1973)
44. Bull, P., Connelly, G.: Body movement and emphasis in speech. J. Nonverbal Behav. 9, 169–187 (1985)

6 Appendix

Speech Linked Gestures (SLG). CD: These gestures are performed during speech exposition. The presence of a concurrent verbal discourse is a necessary but not sufficient condition for the use of such gestures. OD: These gestures have a link, either semantic, referential, or structural, to the speech. The sound, word or verbal utterance to which these gestures are linked, is not strictly synchronic with the gesture: words can slightly precede or follow the concurrent gesture (or be omitted).

Cohesive Gestures. CD: cohesive gestures [9] refer to utterance structure, creating linkages across narrative texts: they are linked with the syntactic aspects of the spoken utterance that determine its structure. OD: cohesive gestures are repetitive similar hand movements [16] performed in the same place and with same shape (each single type having its own idiosyncratic shape, e.g., circular or forward-backward or right-left hand movements). Each sub-category of cohesive gesture is named according to the specific movement shape in the air. For example, "Weaving" (*Matassa* in Italian), in which both hands move horizontally, like if they are weaving.

Rhythmic Gestures. CD: these gestures do not refer to the actual speech content but to prosodic aspects of verbal utterance. OD: they are rhythmical pulsing hand/finger movements (up-down, right-left) in time with co-occurring vocal peak. They are repeated along with the rhythmical pulsation and stress of the speech.

Ideational Gestures. CD: these hand movements are related with the semantic content of the speech they accompany. OD: in such gestures, hands perform movements whose shape explicitly refers to (indicating or representing) concrete or abstract content(s) expressed in concurrent speech.

Emblems Category. [3]. CD: emblems, also called *autonomous gestures* [4], *conventional gestures* [15], *formal pantomimic gestures* [35], *semiotic gestures* [36], are probably the first kind of gestures systematically and scientifically treated, e.g., [37]. They include all symbolic gestures [38], whose specific meaning is widely culturally shared. The same emblem can have different meaning, according to the culture; nevertheless, there are "trans-cultural" hand emblems. OD: Emblems are easily recognizable because, in spite of their arbitrary link with the speech they refer to, they have a direct verbal translation, which would usually consist of one or two words or a whole sentence (often a traditional expression shared in a specific culture). The concurrent words can be completely replaced by an emblematic gesture, as the so called "bag hand" meaning in the Italian culture something like "Well, what do you want from me?" (see [39], for a Sardinian example).

Illustrators. CD: these hand movements, also called *substantive gestures* [15], *topic gestures* [2], *propositional gestures* [9], illustrate the content of what the speaker tells. They enlarge or complete the communication content, indicating something in the space or outlining shapes of objects or movements. Contrary to emblems, link between illustrators shape and meaning is not arbitrary but alludes to some verbal-gesture relation. OD: the shape or the movement drawn by the hand(s) refers to the verbal content (representing or indicating it). To recognize this category is necessary

indentifying its (actual or ideal) referent into the speech [8]. The category of illustrator gestures includes "iconic", "metaphoric", and "deictic" gestures [9].

Iconic gestures reproduce concrete aspects of verbal content. They have a "formal" relation with the referent since their form conveys the meaning and at the same is determined by it. Operatively, hand(s) draw(s), in the air, pictures of objects cited in discourse (e.g., drawing the form of a cube when the speaker mentions "a box").

Metaphoric gestures are also pictorial like iconic ones, but they refer to abstract idea(s), which is concretized through a specific gestural shape. Operatively, the hand, as it moves, "draws", in the air, shapes which can represent a metaphor of an abstract idea, e.g., forming a fist shape when referring to strength: the fist becomes a metaphor of the abstract concept of strength. In an example reported in video on the CodGest, the speaker says "they are trying to make forget the past Governance" and performs a gesture representing a movement to one side, or better "put aside": this movement of putting aside is a metaphor of the abstract action of "forgetting".

Deictic gestures (from ancient Greek language *deìknymi*, "to show"), also named pointing [40], indicate entities which can be actually present in the physic environment of the gesturer (e.g., indicating objects, person, or places) or ideally present in the discourse content (e.g., pointing upwards speaking about northern, or backwards to indicate the past). They can be used for pointing to the interlocutor (as in the case of "interactive" gestures by [2]).

Speech Non-linked Gestures (SNG). CD: According some authors, e.g. [6], these categories could be referred to as hand movement rather than hand gestures. But, referring to Poggi [8, p. 46], gestures are any hand movement performed "to do things, to touch objects, other people or themselves, and finally to communicate". Therefore, SNG can also be produced during speech but do not bear any clear evident relation neither to speech content nor to speech structure (whether in its prosodic or syntactic aspects). OD: these hand movements are mainly acts of contact and/or manipulation with a part of the speaker's body, or with objects, or with other persons, as the adaptors or adapters described by Ekman and Friesen [3] (see also: [41]). SNG are distinguished in "self-adaptor gestures" and "hetero-adaptor gestures". Self-adaptor gestures, also called *body-focused movements* [42], *self-touching gestures* [41, 43], *self-manipulators* ([41]), are gestures of self-contact. Operatively, hands touch parts of one's own body, e.g., touching one's own hair, scratching oneself, rubbing the hands each other. Hetero-adaptors, also called *manipulative gestures* [18], *contact acts* [44], are gestures of contact with what is external to the performer. They can be "person-addressed adaptors" or "object-addressed adaptors". Operatively, object-adaptors are gestures of contact with objects, e.g., touching (manipulating) some object in the physical space such as a pen, or a paper; person-adaptors are contacts with other persons, e.g., touching another one's hand, arm or shoulder.

Individuality in Communicative Bodily Behaviours

Costanza Navarretta

Centre for Language Technology, University of Copenhagen,
Njalsgade 140, build. 25, 4.
2300 Copenhagen S, Denmark
costanza@hum.ku.dk
http://cst.dk/costanza

Abstract. This paper investigates to which extent participants in spontaneously occurring interactions can be recognised automatically from the shape description of their bodily behaviours. For this purpose, we apply classification algorithms to an annotated corpus of Danish dyadic and triadic conversations. The bodily behaviours which we consider are head movement, facial expressions and hand gestures. Although the data used are of limited size, the results of classification are promising especially for hand gestures indicating big variance in people's bodily behaviours even if the involved participants are a homogeneous group in terms of gender, age and social background. The obtained results are not only interesting from a theoretic point of view, but they can also be relevant for video indexing and searching, computer games and other applications which involve multimodal interaction.

Keywords: multimodal annotated spontaneous interactions, classification of participants, individuality.

1 Introduction

This paper investigates to which extent the description of the shape of communicative bodily behaviours annotated in a corpus of spontaneous interactions can be used to automatically individuate their producer (the gesturer, henceforth). Bodily behaviours, which we will simply call gestures henceforth, comprise inter alia gaze, head movements, facial expressions, bodily postures, arm and hand gestures.

Automatically recognising participants in interactions is not only useful in applications which require the identification of people such as video indexing [1], security systems and various interactive systems, but it can also make automatic speaker recognition [2] more robust in multimodal settings.

Furthermore, investigating whether extent various types of communicative gestures are culturally and socially determined or are due to the personality of the subjects producing them is also important from a theoretical point of view.

Most systems for recognizing speakers or persons in multimodal settings focus on speech and facial expressions comprising general face features as well as lips

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 417–423, 2012.

and gaze in particular applications i.a. [3–5]. Differing from these approaches, we do not include speech, but exclusively consider gestures, more precisely head movements, facial expressions and hand gestures, in order to identify the gesturer in naturally occurring conversations.

More precisely, we present a pilot investigation of differences in the shape of gestures produced by various subjects in a subset of a Danish multimodal annotated corpus of spontaneous face-to-face interactions between two or three participants. Because the corpus does not contain any information about the style in which gestures have been performed, such as their intensity or smoothness, we test whether it is possible to individuate the gesturer automatically from the coarse-grained annotations describing the gestures' physical characteristics, the gesture shape henceforth, in the data. The used method is supervised learning. Thus, we depart from preceding work in which we attempted to automatically identify functions of gestures on the basis of their shape independently of their producer, i.a. [6–9].

The paper is organized as follows. First, we present the data (section 2), then we describe our machine learning experiments and discuss the obtained results (section 3). Finally, we conclude and present future work (section 4).

2 The Data

Our data are multimodally annotated video-recordings of two dyadic and two triadic Danish spontaneous interactions between friends and family members [10] for a total duration of approx. 30 minutes. The interactions are recorded in private homes by researchers at University of Southern Denmark [11][1]. The participants in the interactions are five women aged 50+, sitting around a sofa table, drinking coffee and talking about topics such as old days, family relations and the economic crises. The participants, here indicated by the letters A-E, were involved in the interactions in the following way: AB, AC, ADE, CDE.

The interactions have been orthographically transcribed and temporally aligned at the word level and the videos have been multimodally annotated in the ANVIL tool by researchers at University of Copenhagen[2]. The multimodal annotations are based on an extension of the MUMIN annotation model [12] which provides predefined attributes and values for describing the gestures' shape, their communicative functions and their relation to speech. The model deals with the communicative functions of feedback, turn management and sequencing, and focuses on head movements, facial expressions, hand gestures and bodily postures. According to the model, gestures are multifunctional, thus they can be assigned more communicative functions. Gestures can also be given a semiotic type according to the classification of signs proposed by [13]. In our data, a more fine-grained description of the gestures than that proposed in the MUMIN model has been used [6]. The annotation of the shape of head movements

[1] Some of the interactions are available from the talkbank homepage `http://www.talkbank.org`

[2] The annotations were done under the Danish Clarin project.

consists of the description of a) the movement type, such as *Nod*, *SideTurn-Left*, *Tilt*, *HeadBackward*, b) whether the movement is single or repeated, and c) the position of the face with respect to the interlocutor (*FaceTowardInterlocutor*, *FaceAwayFromInterlocutor*). For gaze, information about gaze direction and position with respect to the interlocutor is coded.

In Table 1, the attributes and values describing the shape of hand gestures are shown. These attributes and values are a sub-part of the scheme used at the McNeill Lab [14]. They account for eight dimensions comprising the description of the hand(s) involved in the gesture, the trajectory and amplitude of the movement, as well as the orientation of the palm.

Table 1. Hand gesture description

Behaviour attribute	Behaviour value
Handedness	SingleHand, BothHands
Hand-Repetition	Single, Repeated
Palm	Open, Closed, PalmOther
PalmOrient	Up,Down, Side, Vertical, OrientOther
Fingers	IndexExtended, ThumbExtended,AllExtended,FingersOther
Amplitude	Centre, Periphery,AmplitudeOther
RightHand or LeftHand	Forward, Backward, SideRight, SideLeft, Up, Down, HandComplex, HandOther

We have run an inter-coder agreement experiment in order to test to which extent the various coders recognised the same hand gestures and assigned the same categories to them. In the experiment, two annotators coded hand gestures in a dyadic interaction independently. The results of this experiment in terms of Cohen's kappa [15] are between 0.65-0.85, depending on the categories. These results are satisfactory given the type of task, and are similar to those obtained for other bodily behaviours in other data annotated according to the MUMIN model, inter alia [16, 17].

Table 2 shows the number and type of gesture coded in the dyadic and triadic interactions used in this study. The most frequently occrring gestures in both dyadic and triadic interactions are head movements and gaze.

Table 2. Gestures in dyadic and triadic interactions

Interactions	Face	Head	Gaze	Hand	Body
dyadic	63	346	344	105	51
triadic	47	680	652	263	28
all	110	1026	996	368	69

3 Individuality of Gestures

The MUMIN annotations model was built with the purpose of describing gestures with specific communicative functions. Thus, a number of studies have been conducted on MUMIN annotated data focusing on especially head movements and facial expressions with the communicative functions of feedback, inter alia [18, 9] and turn management [19]. These studies show that there is a certain regularity in the way people use gestures which are related to these functions.

However, it is known that persons make gestures in individual ways, see inter alia [20]. Therefore, in the present study we investigate individual differences in the production of gestures, instead of looking at common uses of gestures related to specific communicative functions. Thus, we investigate to which extent it is possible to identify gesturers automatically from the descriptions of the gestures' shape in our data. This is done applying classification algorithms on the annotated interactions.

Classification is run in WEKA [21] using the SMO algorithm which implements sequential minimal optimisation for training a support vector classifier [22]. The algorithm is evaluated via ten-folds cross-validation, and we use the results of the ZeroR algorithm as baseline [3]. Classification is run on the following three datasets: a) combined head and gaze movements (features: HeadMovement, HeadRepetion, FacetoInterlocutor, GazeDirection, and GazetoInterlocutor), b) facial expressions (features: Face, Eyebrows, Eyes, MouthOpenness, and MouthLips), and c) hand gestures (features: Handedness, Fingers, Palm, PalmOrientation, Amplitude, TrajectoryRightHand, TrajectoryLeftHand,HandRepetition). Then, we run the same experiments on the three data sets enriched with information about the duration of the gestures. In this case, we wanted to test whether the duration of gestures varies significantly from person to person.

The results of the experiments in terms of Precision, Recall and F-score are in Table 3. The F-score is calculated in WEKA as the sum of the weighted F-scores for all class labels. The weight depends on the number of instances in each class. P was sat to 1.0E-12 in all experiments. The results show that the F-score for the classification of the producers of hand gestures is 63.4, while for head movements is 41.4 and for facial expressions is 30.7. In all cases, the results obtained by the SMO algorithm are much better than the baseline.

3.1 Discussion

The results of the classification experiments indicate that looking exclusively at the shape of gestures, especially of hand gestures, can contribute to the automatic identification of the gesturers even if the training data is not large. The results are also promising because the participants in our interactions are quite homogeneous (same gender, same social and age group), thus we can probably expect more variation in the shape of gestures when investigating more subjects especially if they belong to a less homogeneous group than the participants in

[3] ZeroR always chooses the most frequent nominal class in the data.

Table 3. Results of Classification Experiments

Algorithm	Dataset	Precision	Recall	F-score
ZeroR	HeadGaze	6.2	24.9	9.9
SMO	HeadGaze	42.1	43.3	41.4
SMO	HeadGaze+duration	42.1	43.3	41.4
ZeroR	Hand	11.9	34.5	17.7
SMO	Hand	63.6	63.9	63.4
SMO	Hand+duration	63.2	63.3	63
ZeroR	Face	1.19	34.5	17.7
SMO	Face	28.5	38.2	30.7
SMO	Face+duration	28.3	38.2	30.7

our interactions. Since we have applied classification separately to each modality, it can be worth investigating whether the results improve if we consider all gesture types at the same time.

The fact that there are more individual differences in the way people use hand gestures than head movements is not surprising if we consider that hand gestures involve many articulations while head movements do not. Furthermore, many hand gestures are iconic and their shape is per definition not standardised. It might seem surprising that the producers of facial expressions are more difficult to identify in these data than those of head movements because facial expressions are the preferred modality to show emotions and attitudes, and they involve eyebrows, eyes and mouth movements. However, in our data the number of annotated facial expressions is much smaller than that of head movements (see Table 2), and this obviously influences the classification results.

Adding information about the duration of gestures has no influence on their identification. This surprised us because we expected that there would be a relation between the duration of gestures and their intensity, which is one of the parameters recognised in the literature as personality indicators [23]. One explanation of the missing relation between duration of gestures and gesturers in these data can be that subjects in these interactions knew each other very well, thus they are all relaxed while conversing, and they might somehow coordinate the rhythm of their verbal and not verbal behaviours. This aspect, however, must be tested on more data, and the behaviours of the same subjects in different types of interactions and with varying interlocutors should be compared.

4 Concluding Remarks and Future Work

In this paper, we have described machine learning experiments on annotated multimodal spontaneous interactions between Danish well-acquainted people. These experiments were run in order to investigate how far the producers of head movements, facial expressions and hand gestures can be automatically identified on the basis of coarse-grained descriptions of the shape of these gestures. The

results of a support vector classifier are much better than the baseline for all gesture types, but they also show that the producers of hand gestures can be identified with greater accuracy than the producers of head movements and facial expressions. Adding information about the duration of gestures did not influence classification at all.

Although the data we have used have a limited size, the obtained results are promising because the population is very homogeneous. Since we have run classification on each modality independently, we expect that combining the various modalities will give even better results. Recognising automatically the participants in interactions from their gestures can be useful in applications such as video indexing, video searching and interactive systems including computer games. Our results can also contribute to studies on gesture production, because they show that the shape of gestures, especially hand gestures, varies greatly from subject to subject even when they have the same gender and belong to the same age and social group.

Future work will consist in running similar experiments on more data and testing some of the hypothesis proposed in the paper in more conversational settings with a more heterogeneous participant population. We will also investigate individual differences in the production of verbal and bodily behaviours combining the various modalities, instead of taking them as independent events.

Acknowledgements. The present work was done under the VKK project funded by the Danish Research Councils. Thanks to Patrizia Paggio, Jens Allwood, Elisabeth Ahlsén and Kristiina Jokinen for the many useful discussions on multimodal communication and annotation of multimodal data.

References

1. Idris, F., Panchanathan, S.: Review of Image and Video Indexing Techniques. J. Vis. Commun. Image R 8(2), 146–166 (1997)
2. Beigi, S.: Fundamentals of Speaker Recognition. Springer (2011)
3. Erzin, E., Yemez, Y., Tekalp, A.M.: Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Reliability. IEEE Trans. Multimedia 7(5), 840–852 (2005)
4. Wu, Z., Cai, L., Meng, H.: Multi-level fusion of audio and visual features for speaker identification. In: International Conference on Advances in Biometrics, pp. 493–499 (2006)
5. Wee-Chung, A., Wang, L., Wang, S. (ed.): Speech Recognition: Lip Segmentation and Mapping. IGI Global (2009)
6. Navarretta, C.: Anaphora and gestures in multimodal communication. In: Hendrickx, Branco, Devi, L., Mitkov (eds.) Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Faro, Portugal, Edicoes Colibri, 171–181 (2011)
7. Paggio, P., Navarretta, C.: Feedback in Head Gestures and Speech. In: Kipp, M., et al. (eds.) LREC 2010 Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pp. 1–4 (2010)

8. Navarretta, C., Paggio, P.: Classification of Feedback Expressions in Multimodal Data. In: Proceedings of ACL 2010, Uppsala, Sweden, pp. 318–324 (2010)
9. Paggio, P., Navarretta, C.: C. Learning to classify the feedback function of head movements in a Danish Corpus of first encounters. In: Proceedings of ICMI 2011 Workshop Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future, Alicante, Spain, pp. 49–54 (November 2011)
10. Navarretta, C.: Annotating Non-verbal Behaviours in Informal Interactions. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 309–315. Springer, Heidelberg (2011)
11. MacWhinney, B., Wagner, J.: Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. Gespraechsforschung 11, 154–173 (2010)
12. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In: Martin, J.-C., et al. (eds.) Multimodal Corpora for Modelling Human Multimodal Behaviour, Special issue of the International JLRE, pp. 273–287. Springer (2007)
13. Peirce, C.S.: Collected Papers of Charles Sanders Peirce, 1931-1958, 8 vols, Hartshorne, C., Weiss, P., Burks, A. (eds.). Harvard University Press, Cambridge (1931)
14. Duncan, S.: McNeill Lab Coding Methods. Technical Report (2004)
15. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20(1), 37–46 (1960)
16. Jokinen, K., Navarretta, C., Paggio, P.: Distinguishing the Communicative Functions of Gestures. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 38–49. Springer, Heidelberg (2008)
17. Navarretta, C., Ahlseen, E., Allwood, J., Jokinen, K., Paggio, P.: Creating Comparable Multimodal Corpora for Nordic Languages. In: Proceedings of Nodalida 2011, pp. 153–160 (2011)
18. Lu, J., Allwood, J., Ahlsén, E.: A Study on Cultural Variations of Smile Based on Empirical Recordings of Chinese and Swedish First Encounters. In: Proceedings of ICMI 2011 Workshop Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future, Alicante, Spain, pp. 37–42 (November 2011)
19. Jokinen, K.: Turn taking, Utterance Density, and Gaze Patterns as Cues to Conversational Activity. In: Proceedings of ICMI 2011 Workshop Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future, Alicante, Spain, pp. 31–36 (November 2011)
20. Kipp, M.: Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. Ph.D. thesis, Saarland University, Saarbruecken, Germany, Boca Raton, Florida, dissertation.com (2004)
21. Witten, J.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
22. Platt, J.C.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning, pp. 41–65. MIT Press (1998)
23. Gallaher, P.E.: Individual differences in nonverbal behavior: Dimensions of style. J. Pers. Soc. Psychol. 63(1), 133–145 (1992)

A Cross-Cultural Study on the Perception of Emotions: How Hungarian Subjects Evaluate American and Italian Emotional Expressions

Maria Teresa Riviello[1], Anna Esposito[1], and Klara Vicsi[2]

[1] Seconda Università degli Studi di Napoli, Department of Psychology, and IIASS, Italy
[2] Laboratory of Speech Acoustics, Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, Budapest, Hungary
mariateresa.riviello@unina2.it, iiass.annaesp@tin.it, vicsi@tmit.bme.hu

Abstract. In the present work a cross-modal evaluation of the visual and auditory channels in conveying emotional information is conducted through perceptual experiments aimed at investigating whether some of the basic emotions are perceptually privileged and whether the perceptual mode, the cultural environment and the language play a role in this preference. To this aim, Hungarian subjects were requested to assess emotional stimuli extracted from Italian and American movies in the single (either mute video or audio alone) and combined audio-video mode. Results showed that among the proposed emotions, anger plays a special role and fear, happiness and sadness are better perceived than surprise and irony in both the cultural environments. The perception of emotions is affected by the communication mode and the language influences the perceptual assessment of emotional information.

Keywords: ultimodal Expression of Emotion, Perceptually Privileged Emotions, Cultural and Language Specificity Effect.

1 Introduction

The role of emotion in the context of Human Computer Interaction (HCI) is becoming ever more relevant and challenging. HCI for affective systems embraces theories from a wide range of domains and disciplines such as psychology and sociology, robotics, computer science or design. It is relevant to a diverse set of application areas, from teaching and learning to office applications, entertainment technology, therapeutic applications, through to advertising and product design. Continual areas of interest within research include the recognition as well as the synthesis of affect and emotion in the face, body and speech. Given the complexity and the multimodal nature of the phenomenon, there has been a branching of the engineering approach toward the improvement and development of video-audio processing, recognition and synthesis techniques [3-4, 7-8, 12, 17-19, 29, 30-31] with the goal to develop new methodologies for recognizing emotional states from faces [26, 33-37], speech [1-2, 20-24, 43, 45] and/or body movements [5, 42].

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 424–433, 2012.

Since all the above research lines imply the analysis of verbal and nonverbal behaviors in human interactions, further efforts are worth making in order to go into the details of the decoding process, investigating the precise cues the addressees use in inferring the addressers' emotional state.

The present research is focused on the cross modal analysis of realistic emotional stimuli in the attempt to clarify the mechanisms underlying the human perception of emotional expressions, as well as for identifying cross-cultural differences among such perceptual processes.

In the present work, Hungarian subjects were involved in perceptual experiments aimed at investigating their ability in identifying emotional expressions extracted from American English and Italian movies, and dynamically presented through the visual and auditory channels, considered either singularly (either mute video or audio alone) or in combination (combined audio-video mode).

The experimental set up proposed allows to explore the possible effect of culture and language specificity on the emotional information decoding process, investigating whether the familiarity of the language and the subjects' exposition to the cultural environment affect their recognition of the emotional stimuli especially when they are transmitted through the vocal channel [15, 40]. In this particular case, it explores the amount of emotional content that Hungarian participants are able to infer from emotional stimuli belonging to two different languages (American English, a West Germanic language considered here as a global spread language and Italian, a Romance language, considered here as a country specific language), far from their own (Hungarian is a Ugrian subgroup of Uralic languages and is not part of the Indo-European family tree).

2 Materials

Two databases of emotional stimuli, which allow to compare dynamic visual and vocal information, were defined exploiting video-clips extracted from American and Italian movies. The use of audio and video stimuli extracted from movies provided a set of realistic emotional expressions [13-16]. Differently from the other existing emotional databases proposed in literature [9, 32, 38], in this case the actors/actresses had not been asked to produce an emotional expression, rather, they were acting according to a movie script and their performance had been judged as appropriate to the required emotional context by the movie director (supposed to be an expert).

Each database consists of audio and video stimuli representing 6 emotional states: *happiness*, *sarcasm/irony*, *fear*, *anger*, *surprise,* and *sadness* (except for sarcasm/irony, the remaining emotions are considered as basic and therefore universally shared [4, 10, 27, and 39]). For each database and for each of the above emotional states, 10 stimuli were identified, 5 expressed by an actor and 5 by an actress, for a total of 60 American and 60 Italian video-clips, each acted by a different actor and actress to avoid bias in their ability to portray emotional states. The stimuli are short in duration (the average length was 3s, SD = ± 1s) to avoid the overlapping of emotional states that could confuse the subject's perception. In the selected video-clips the protagonist's face and the upper part of the body are clearly visible. In addition, the semantic meaning of the produced utterances is not clearly expressing the portrayed emotional state and its intensity level is moderate.

The emotional labels assigned to the stimuli were first given by two experts and then by three naïve judges independently. The expert judges labeled the stimuli carefully exploiting emotional information on facial and vocal expressions such as frame by frame analysis of changes in facial muscles and F0 contour, rising and falling of intonation contour, etc, as reported by several authors in literature [9, 41, and 44], and the contextual situation the protagonist was interpreting. The naïve judges made their decision after watching the stimuli several times. There were no opinion exchanges between the experts and naïve judges and the final agreement on the labeling between the two groups was 100%.

The collected stimuli, being extracted from movie scenes contain environmental noise and therefore are also useful for testing realistic computer applications. The database is available in the context of the COST Action 2102 (cost2102.cs.stir.ac.uk/) and can be required by mailing the chair of COST 2102. Both for the American and Italian database, the audio and mute video were extracted from each complete audio-video stimulus (video-clip) coming up with a total of 180 American and 180 Italian stimuli: 60 audio, 60 mute video and 60 combined audio-video stimuli for each database. The stimuli in each set were then randomized and proposed to the subjects participating at the experiments.

3 Participants

A total of 180 Hungarian subjects (90 for the Italian and 90 for the American database) participated at the perceptual experiments. For each database, 30 subjects were involved in the evaluation of the audio stimuli, 30 in the evaluation of the mute video stimuli and 30 in the evaluation of the audio-video stimuli. In each group 15 participants were male and 15 were female, aged from 18 to 35 years. All of them used English as a second language, whereas they had no knowledge of Italian.

The subjects were randomly assigned to the task and were required to carefully listen to and/or watch the stimuli via headphones in a quiet room. They were instructed to pay attention to each presentation and decide which of the 6 emotional states was expressed in it. Responses were recorded on matrix paper form (60x8), where rows listed the stimuli numbers and columns the 6 selected emotional states plus "*others*" indicating any other emotion not listed, and the option of "*no emotion"* that was suggested when according to the subject's feeling the protagonist was showing no emotion.

For each emotional stimulus, both the frequency response distribution among the 8 emotional classes under examination (*happiness*, *sarcasm/irony*, *fear*, *anger*, *surprise*, *sadness, others,* and *no emotion*) and the percentage of correct recognition were computed.

4 Hungarian Subjects Tested on American Emotional Stimuli

The results obtained by the Hungarian subjects assessing emotional stimuli extracted from American movies are summarized in Figure 1, where it is displayed the

percentage of label's agreement obtained under the three different experimental conditions (audio alone, mute video and combined audio-video).

The data essentially show that Hungarian subjects did extract emotional information from both the auditory and visual channels, even though they rely more on visual information.

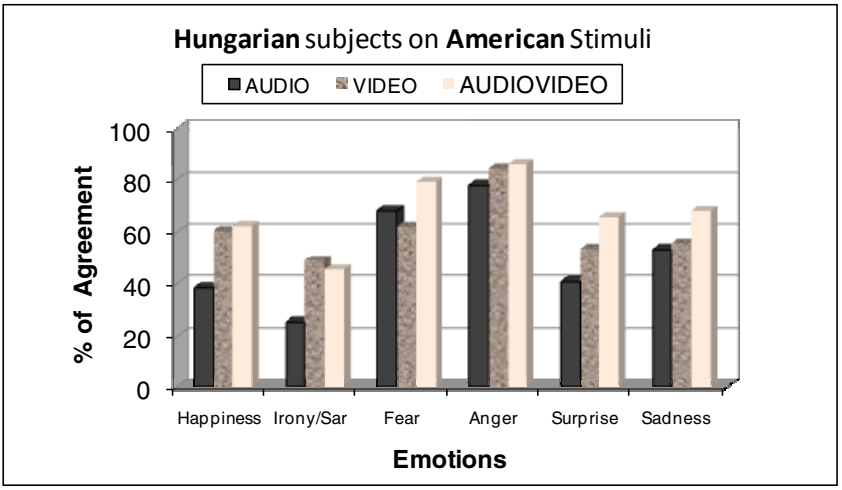


Fig. 1. Percentage of agreement obtained under the three experimental conditions (only audio – black bars- only mute video –gray bars- audio-video-white bar) by the Hungarian subjects tested on American stimuli

An ANOVA analysis was performed on the data obtained by the Hungarian subjects tested on American stimuli, considering the *Perceptual mode* as a between subjects variable and the *Emotions and Actor's Gender* as within subjects variables. Significance was established for α=.05.

The ANOVA shows that the perceptual mode plays a significant role ($F(2, 12) = 10.455$, $\rho = .002$). The recognition of a given emotional states significantly depends on the portrayed emotion ($F(5, 60) = 5.077$, $\rho = .0006$), independently from the perceptual mode ($F(10, 60) = .951$, $\rho = .50$). Emotion recognition accuracy is also affected by the gender of the protagonist portraying the emotional expressions ($F(1, 12) = 69.797$, $\rho = .0001$), independently from the perceptual mode, since no interaction was found between the two variables ($F(2, 12) = 2.388$ $\rho = .13$). An interaction was found between the emotion category and the actor's gender ($F(5, 60) = 5.475$, $\rho = .0003$).

Hungarian subjects tested on American stimuli rely more on visual information: mute video and combined audio-video ($F(1, 8) = 2.749$, $\rho = .13$) convey the same amount of emotional information, while there are significant differences between audio alone and combined audio-video ($F(1, 8) = 29.276$, $\rho = .0006$), and audio alone and mute video ($F(1, 8) = 6.789$, $\rho = .03$).

.

5 Hungarian Subjects Tested on Italian Emotional Stimuli

Figure 2 reports for each emotion the percentage of agreement expressed by Hungarian subjects participating to the experiments for the Italian emotional stimuli. As in Figures 1, on the x-axis are the emotions under consideration and on the y-axis is reported (for each emotion) the percentage of correct agreement under the three experimental conditions.

The data revealed a stronger preference of the Hungarian subjects for the visual channel when tested on Italian stimuli than on the American ones.

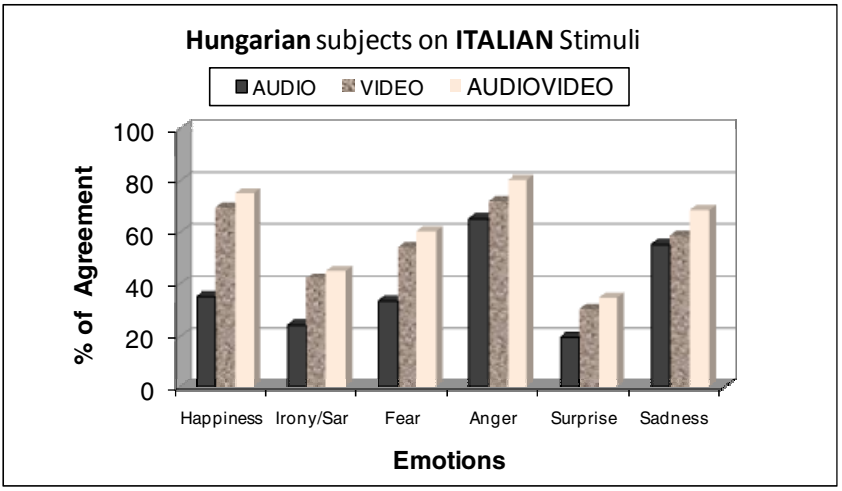


Fig. 2. Percentage of agreement obtained under the three experimental conditions (only audio – blackbars- only mute video –gray bars-combined audio-video-white bar) by the Hungarian subjects tested on Italian stimuli

An ANOVA analysis as described above was performed on the frequency of correct answers obtained by the Hungarian subjects when tested on Italian stimuli. As expected the perceptual mode plays a significant role ($F(2, 12) = 9.223$, $\rho = .004$).

The differences are not significant between the mute video and the audio-video condition ($F(1, 8) = .996$, $\rho = .35$), whereas they are significant between audio alone and combined audio-video ($F(1, 8) = 23.761$, $\rho = .001$), and audio alone and mute video ($F(1, 8) = 10.145$, $\rho = .01$).

Identification among emotions is significantly different ($F(5, 60) = 6.638$, $\rho = .0001$), independently from the perceptual mode ($F(10, 60) = .948$, $\rho = .50$). Differently from the data for the American stimuli, the gender of the protagonist doesn't affect the recognition accuracy ($F(1, 12) = 1.510$, $\rho = .25$) and no interaction was found between the perceptual mode and the actors' gender ($F(2, 12) = .766$ $\rho = .49$). However, an interaction was found between the category of emotion and the actor's gender ($F(5, 60) = 12.363$, $\rho = .00001$).

6 Assessment of the Performance among the Two Cultures

To assess the role of language and the cultural context characterizing the emotional expressions used as stimuli, the Hungarian recognition accuracy for the American and Italian audio, the mute video and the combined audio-video are reported in Figures 3, 4 and 5 respectively.

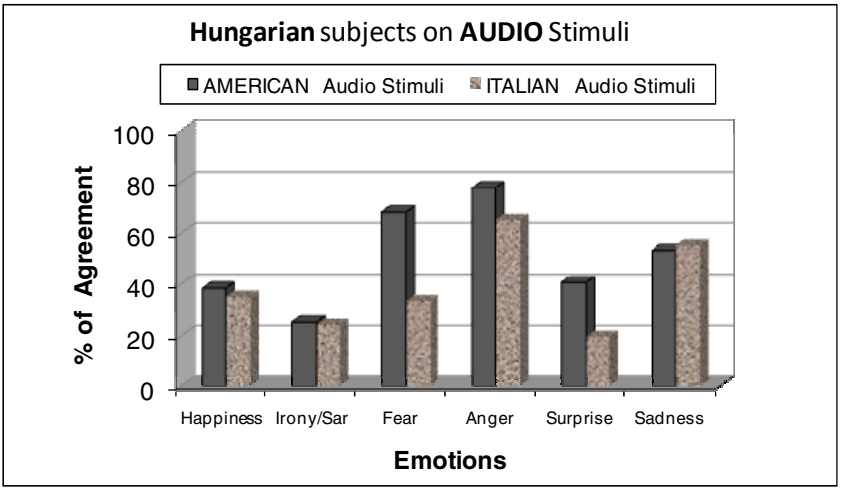


Fig. 3. Percentage of agreement obtained by the Hungarian subjects tested on American (black bar) and Italian (gray bar) audio stimuli

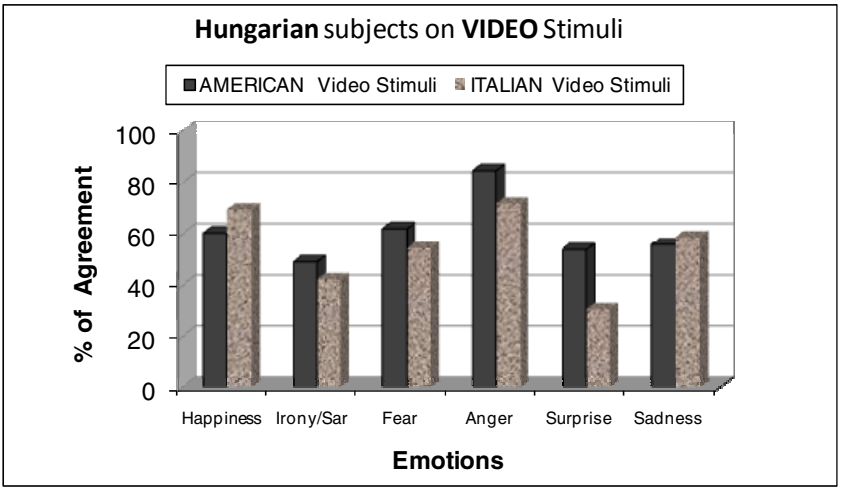


Fig. 4. Percentage of agreement obtained by the Hungarian subjects tested on American (black bar) and Italian (gray bar) mute video stimuli

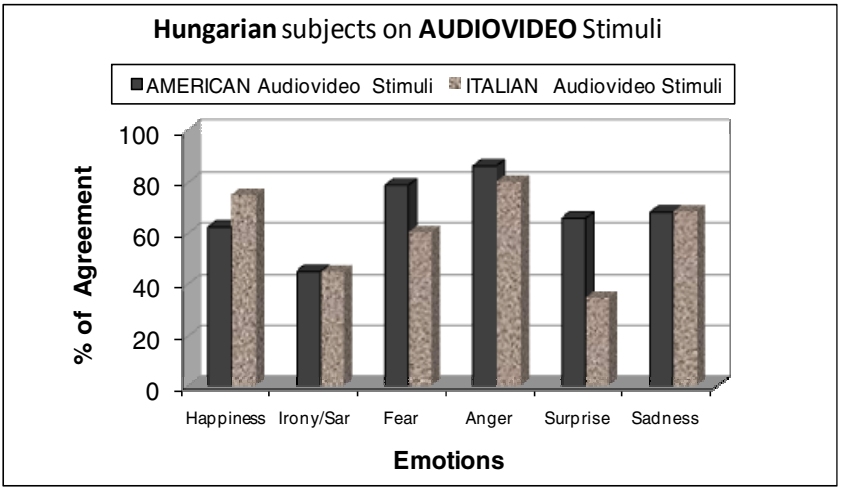


Fig. 5. Percentage of agreement obtained by the Hungarian subjects tested on American (black bar) and Italian (gray bar) combined audio-video stimuli

In this case, the ANOVA analyses were separately performed on the data obtained in the three different perceptual modes (i.e., the audio alone, the mute video and the combined audio-video), considering the *Stimulus Language/Culture* (American and Italian) as a between subjects variable and the *Emotions* and *Actors' Gender* as within subjects variables. Significance was established for α=.05.

The analyses revealed that for the audio alone condition the stimulus language/culture significantly affects the subjects' recognition accuracy (F (1, 8) = 18,892, ρ=.002). In particular the American emotional vocal expressions were better recognized than the Italian ones (American mean values=15,167, SD=.575; Italian mean values=11,633 SD=.575). No significant differences were found in the identification of the mute American and Italian video (F (1, 8) =1,248, ρ=.292) and the combined audio–video (F (1, 8) = 2,248, ρ=.172) respectively.

7 Discussion and Conclusions

In this study Hungarian subjects were involved in perceptual experiments devoted to assess their ability to identify emotional expressions through visual and auditory channels, considered either singularly or in combination, exploiting two databases of stimuli based on video-clips extracted from American and Italian movies, respectively, and representing six primary emotional states: *happiness*, *sarcasm/irony*, *fear*, *anger*, *surprise* and *sadness.*

The results show that among the primary emotions [6, 28] anger is perceptually privileged since regardless of the cultural context and the communication mode, it is always better recognized than the others.

Assuming that the experimental set up proposed is testing the capability of the subjects to infer emotional cues from the interlocutor, identifying anger may activate cognitive self-defense mechanisms that are crucial for the perceiver's survival, hence

humans may have a high sensitivity to recognize it independently from the cultural environment.

Among the remaining emotional states under consideration, sadness is equally well recognized by the Hungarian subjects in both the cultural environments, whereas fear and surprise are better identified for the American stimuli than for the Italian ones, while happiness cues are better perceived for the Italian visual stimuli than for the American ones. The Hungarian subjects showed difficulties to recognize both the American and Italian ironic expressions, probably because the perceptual cues of this emotion are strictly linked to the language and the original cultural context.

The above data also show that in perceiving emotions, expressive information is not added over the amount of emotional cues provided, corroborating the idea of a nonlinear processing of emotional cues [13-16]. In particular, the audio and video combined conveys the same amount of emotional information conveyed by the mute video. It can be hypothesized that in the multimodal presentation, the subjects perception is affected by a cognitive load caused by the concurrent processing of dynamically evolving visual and vocal expressions and by prosodic emotional features of the produced speech [13]. Subjects attempt to reduce this cognitive load taking advantage of their ability to process visual stimuli especially when the auditory input is in a foreign language, probably because the assessment of visual cues involves processing mechanisms similar across cultures whereas the processing of speech inputs requires a specific language expertise. This hypothesis can also account for the Hungarian greater difficulty in identifying Italian emotional vocal expressions since this language is culturally far and completely unknown with respect to the American language, globally spread and used as second language by the involved subjects.

Acknowledgements. This work has been supported by the European projects: COST 2102 “Cross Modal Analysis of Verbal and Nonverbal Communication”, http://cost2102.cs.stir.ac.uk/ and COST ISCH TD0904 “TMELY: Time in MEntal activity (http://w3.cost.eu/index.php?id=233&action_number=TD0904). Acknowledgements go to Miss Tina Marcella Nappi for her editorial help.

References

1. Apolloni, B., Aversano, G., Esposito, A.: Preprocessing and classification of emotional features in speech sentences. In: Kotare, Y. (ed.) Proceedings of the IEEE Workshop on Speech and Computer, pp. 49–52 (2000)
2. Apolloni, B., Esposito, A., Malchiodi, D., Orovas, C., Palmas, G., Taylor, J.G.: A general framework for learning rules from data. IEEE Transactions on Neural Networks 15(6), 1333–1350 (2004)
3. Apple, W., Hecht, K.: Speaking emotionally: The relation between verbal and vocal communication of affect. Journal of Personality and Social Psychology 42, 864–875 (1982)
4. Banse, R., Scherer, K.: Acoustic profiles in vocal emotion expression. Journal of Personality & Social Psychology 70(3), 614–636 (1996)
5. Bryll, R., Quek, F., Esposito, A.: Automatic hand hold detection in natural conversation. In: Proceedings of IEEE Workshop on Cues in Communication, Hawai, December 9 (2001)

6. Campos, J.J., Barrett, K., Lamb, M.E., Goldsmith, H.H., Stenberg, C.: Socioemotional development. In: Haith, M.M., Campos, J.J. (eds.) Handbook of Child Psychology, 4th edn., vol. 2, pp. 783–915. Wiley, New York (1983)
7. Chollet, G., Esposito, A., Gentes, A., Horain, P., Karam, W., Li, Z., Pelachaud, C., Perrot, P., Petrovska-Delacrétaz, D., Zhou, D., Zouari, L.: Multimodal Human Machine Interactions in Virtual and Augmented Reality. In: Esposito, A., Hussain, A., Marinaro, M., Martone, R. (eds.) COST Action 2102. LNCS, vol. 5398, pp. 1–23. Springer, Heidelberg (2009)
8. Doyle, P.: When is a communicative agent a good idea? In: Proceedings of Inter. Workshop on Communicative and Autonomous Agents, Seattle (1999)
9. Ekman, P., Friesen, W.V., Hager, J.C.: The facial action coding system: Research Nexus eBook, 2nd edn. Weidenfeld & Nicolson, London (2002)
10. Ekman, P.: An argument for basic emotions. Cognition and Emotion 6, 169–200 (1992)
11. Ekman, P.: The argument and evidence about universals in facial expressions of emotion. In: Wagner, H., Manstead, A. (eds.) Handbook of Social Psychophysiology, pp. 143–164. Wiley, Chichester (1989)
12. Elliott, C.D.: The affective reasoned: A process model of emotion in a multi-agent system. Ph.D Thesis, Institute for Learning sciences, Northwestern University, Evanston, Illinois (1992)
13. Esposito, A.: The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information. Cognitive Computation Journal 1(2), 268–278 (2009)
14. Esposito, A., Riviello, M.T., Di Maio, G.: The COST 2102 Italian Audio and Video Emotional Database. In: Apolloni, B., et al. (eds.) Frontiers in Artificial Intelligence and Applications, vol. 204, pp. 51–61 (2009) ISBN 978-1-60750-072-8 (print) ISBN 978-1-60750-515-0, http://www.booksonline.iospress.nl/Content/View.aspx?piid=14188
15. Esposito, A., Riviello, M.T., Bourbakis, N.: Cultural Specific Effects on the Recognition of Basic Emotions: A Study on Italian Subjects. In: Holzinger, A., Miesenberger, K. (eds.) USAB 2009. LNCS, vol. 5889, pp. 135–148. Springer, Heidelberg (2009)
16. Esposito, A.: The Amount of Information on Emotional States Conveyed by the Verbal and Nonverbal Channels: Some Perceptual Data. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) WNSP 2005. 277. LNCS, vol. 4391, pp. 249–268. Springer, Heidelberg (2007)
17. Ezzat, T., Geiger, G., Poggio, T.: Trainable video-realistic speech animation. In: Proceedings of SIGGRAPH, San Antonio, Texas, pp. 388–397 (July 2002)
18. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition 36, 259–275 (2002)
19. Frick, R.: Communicating emotions: the role of prosodic features. Psychological Bullettin 93, 412–429 (1985)
20. Friend, M.: Developmental changes in sensitivity to vocal paralanguage. Developmental Science 3, 148–162 (2000)
21. Fulcher, J.A.: Vocal affect expression as an indicator of affective response. Behavior Research Methods, Instruments, & Computers 23, 306–313 (1991)
22. Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P., Garcia, O.N.: Audio/visual mapping with cross-modal Hidden Markov Models. IEEE Transactions on Multimedia 7(2), 243–252 (2005)
23. Gutierrez-Osuna, R., Kakumanu, P., Esposito, A., Garcia, O.N., Bojorquez, A., Castello, J., Rudomin, I.: Speech-driven facial animation with realistic dynamic. IEEE Transactions on Multimedia 7(1), 33–42 (2005)
24. Hozjan, V., Kacic, Z.: A rule-based emotion-dependent feature extraction method for emotion analysis from speech. Journal of the Acoustical Society of America 119(5), 3109–3120 (2006)

25. Hozjan, V., Kacic, Z.: Context-independent multilingual emotion recognition from speech signals. International Journal of Speech Technology 6, 311–320 (2003)
26. Huang, C.L., Huang, Y.M.: Facial expression recognition using model-based feature extraction and action parameters Classification. Journal of Visual Communication and Image Representation 8(3), 278–290 (1997)
27. Izard, C.E.: Innate and universal facial expressions: Evidence from developmental and cross-cultural research. Psychological Bulletin 115, 288–299 (1994)
28. Izard, C.E.: Human Emotions. Plenum Press, New York (1977)
29. Kähler, K., Haber, J., Seidel, H.: Geometry-based muscle modeling for facial animation. In: Proceedings of the International Conference on Graphics Interface, pp. 27–36 (2001)
30. Kakumanu, P., Esposito, A., Garcia, O.N., Gutierrez-Osuna, R.: A comparison of acoustic coding models for speech-driven facial animation. Speech Communication 48, 598–615 (2006)
31. Kakumanu, P., Gutierrez-Osuna, R., Esposito, A., Bryll, R., Goshtasby, A., Garcia, O.N.: Speech Dirven Facial Animation. In: Proceedings of ACM Workshop on Perceptive User Interfaces, Orlando, November 15-16 (2001)
32. Kamachi, M., Lyons, M., Gyoba, J.: Japanese Female Facial Expression Database, Psychology Department in Kyushu University, http://www.kasrl.org/jaffe.html
33. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
34. Koda, T.: Agents with faces: A study on the effect of personification of software agents. Master Thesis, MIT Media Lab, Cambridge (1996)
35. Morishima, S.: Face analysis and synthesis. IEEE Signal Processing Magazine 18(3), 26–34 (2001)
36. Pantic, M., Patras, I., Rothkrantz, J.M.: Facial action recognition in face profile image sequences. In: Proceedings IEEE International Conference Multimedia and Expo., pp. 37–40 (2002)
37. Pantic, M., Rothkrantz, J.M.: Expert system for automatic analysis of facial expression. Image and Vision Computing Journal 18(11), 881–905 (2000)
38. Samaria, F., Harter A.: The ORL Database of Faces. Cambridge University Press, Cambridge, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
39. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256 (2003)
40. Scherer, K.R., Banse, R., Wallbott, H.G.: Emotion inferences from vocal expression correlate across languages and cultures. Journal of Cross-Cultural Psychology 32, 76–92 (2001)
41. Scherer, R., Oshinsky, J.S.: Cue utilization in emotion attribution from auditory stimuli. Motivation and Emotion 1, 331–346 (1977)
42. Stocky, T., Cassell, J.: Shared reality: Spatial intelligence in intuitive user interfaces. In: Proceedings of Intelligent User Interfaces, San Francisco, CA, pp. 224–225 (2002)
43. Tóth, S.L., Sztahó, D., Vicsi, K.: Speech Emotion Perception by Human and Machine. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction 2007. LNCS (LNAI), vol. 5042, pp. 213–224. Springer, Heidelberg (2008)
44. Ververidis, D., Kotropoulos, C.: Emotional Speech Recognition: Resources, Features and Methods. Elsevier Speech Communication 48(9), 1162–1181 (2006)
45. Vicsi, K., Szaszák, G.: Using prosody to improve automatic speech recognition. Speech Communication 52(5), 413–426 (2010)

Affective Computing: A Reverence for a Century of Research

Egon L. van den Broek

TNO Technical Sciences
P.O. Box 5050, 2600 GB Delft, The Netherlands
Human Media Interaction (HMI), Faculty of EEMCS, University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
Karakter University Center, Radboud University Medical Center Nijmegen
P.O. Box 9101, 6500 HB Nijmegen, The Netherlands
vandenbroek@acm.org

Abstract. To bring affective computing a leap forward, it is best to start with a step back. A century of research has been conducted on topics, which are crucial for affective computing. Understanding this vast amount of research will accelerate progress on affective computing. Therefore, this article provides an overview of the history of affective computing. The complexity of affect will be described by discussing i) the relation between body and mind, ii) cognitive processes (i.e., attention, memory, and decision making), and iii) affective computing's I/O. Subsequently, definitions are provided of affect and related constructs (i.e., emotion, mood, interpersonal stances, attitude, and personality traits) and of affective computing. Perhaps when these elements are embraced by the community of affective computing, it will us a step closer in bridging its semantic gap.

Keywords: affect, emotion, affective computing, biosignals, history, complexity, definitions, semantic gap.

"The increasing tendency to 'let's see what the computer shows' has dulled many investigators' sensitivity to the basic rules of research, sometimes to the point that the definition of the problem is unclear. The speed and ease of statistical computation, even of complex multivariate analysis involving N*s of 1000 or more, have tempted many investigators to submit all possible comparisons for analysis with little concern over the subtleties of multiple testing and the risk of type I errors. The computer cannot substitute for thoughtful planning or, as Medawar (1969, p. 29) put it,* 'We cannot browse over the field of nature like cows at pasture' *"* (Chapter 6, p. 333) [41]

1 Introduction

Almost half a century ago, the American psychologist Ulrich [47] described "*three fundamental and interrelated characteristics of human thought that are conspicuously absent from existing or contemplated computer programs*:

A. Esposito et al. (Eds.): Cognitive Behavioural Systems 2011, LNCS 7403, pp. 434–448, 2012.

1. *Human thinking always takes place in, and contributes to, a cumulative process of growth and development.*
2. *Human thinking begins in an intimate association with emotions and feelings which is never entirely lost.*
3. *Almost all human activity, including thinking, serves not one but a multiplicity of motives at the same time.*" (p. 194)

Ulrich [47] was not only one with this opinion, Nobel price winner and recipient of the ACM's Turing Award, Herbert A. Simon had similar ideas on this topic. He showed "*. . . how motivational and emotional controls over cognition can be incorporated into an information-processing system , so that thinking will take place in 'intimate association with emotions and feelings,' and will serve a 'multiplicity of motives at the same time.'*" (p. 29) [64]. Nonetheless, in the decades that followed Artificial Intelligence (AI) aimed at understanding human cognition *without* taking emotion into account [56].

Although emotions were sometimes denoted as important (e.g., [43, 44]), it took until the publication of Picard's book *Affective computing* [50] before they received center stage attention. Even though AI has made it possible that a computer can beat the world's best chess players [20] and can win quizzes such as Jeopardy! [28], the general opinion is that AI has failed [40] (cf. [23, 36]). This is likely to be (partly) because of a lack of focus on emotions. So, 50 years after Ulric Neisser's words, with the user more demanding than ever, perhaps now is the time to bring emotions to the front line of AI research and practice.

Affective computing includes signal processing and machine learning techniques [12, 14, 16], which can rely on a thorough foundation. Hence, this are not the aspects of *affective computing* that slow down its progress. I pose that the true complexity lies in i) the definition of constructs related to *affective computing*, ii) their operationalization, and, subsequently, iii) their mapping upon the biosignals (or other signals). Over a century of research has been conducted on these three issues, which are crucial for *affective computing*. Therefore, I pose that to bring *affective computing* a leap forward, it is best to start with a step back. Understanding this vast amount of research will accelerate progress on *affective computing*. Consequently, this article provides an overview of the history of *affective computing* (Section 2). Next, in Section 3, the complexity of affect will be discussed. Section 4 provides definitions for emotion, affect, affective computing, and related notions. In Section 5, I end this article with a discussion.

2 Historical Reflection[1]

Our knowledge on emotions has increased significantly over the last centuries [2]. However, often this knowledge is ignored to a great extent and the same (relatively) recent theories are adopted (e.g., the valence-arousal model). This phenomenon is in particular present in the field of *affective computing*, where

[1] This section is based on Section 2.1 of Part IV of [1].

an engineering is often valued more than a solid theoretical framework [1]. So comes that in practice most engineering approaches embrace the valence-arousal model as being the standard, without considering other models and/or theories. Nevertheless, an increase in awareness of other theories could heighten the understanding and, subsequently, contribute to the success of *affective computing*.

A vast number of handbooks and review papers on emotions, affective sciences, and affective neuroscience have appeared over the last 50 years. These include [25, 26, 39, 57, 59], which can all be considered as essential reading material. Regrettably. it is far beyond the scope of this chapter article to provide an exhaustive literature survey on *affective computing*. Therefore, this section will only touch upon some of the main works on emotion research, which originate from biology, medicine, physiology, and psychology.

In 1780, a book was published that presented the work of M. l'Abbé Bertholon with as its title *De l'Électricité du corps humain* [7]. This book is one of the earliest works that described human biosignals. In 1872, nearly a century later, *The expression of emotions in man and animals* of Darwin was published [25]. Next, independently of each other, William James and C.G. Lange coined their theories on emotions. These two theories showed to be remarkably similar and, hence, were merged and were baptized the James-Lange theory [25].

In sum, the James-Lange theory poses that the percept of our biosignals *are* what we denote as emotions. This implies that emotions cannot be experienced without these biosignals. This position was seriously challenged by both Cannon [21, 22] and Bard [3, 4]. Both Cannon and Bard denoted the important role of subcortical structures (e.g., the thalamus, the hypothalamus, and the amygdala) in our experience of emotions. Their theory was founded on five notions:

1. Emotions are experienced similar both with and without (e.g., as with the transection of the spinal cord and vagus nerve) biosignals.
2. Similar biosignals emerge with multiple emotions. Hence, these signals cannot cause these distinct emotions.
3. People's internal organs have fewer sensory nerves than other anatomical structures. This causes people to be unaware of their possible biosignals up to a high extent.
4. Most often, biosignals have a long latency compared to the duration of the emotional responses.
5. Drugs that set off biosignals do not inevitably set off emotions in simultaneously.

Next, I will discuss these five notions in light of *affective computing*. This is of importance for *affective computing* as will become apparent.

In 1966, Hohnmann [30] described a patient who reported: *"Sometimes I act angry when I see some injustice. I yell and cuss and raise hell, because if you don't do it sometimes, I learned people will take advantage of you, but it just doesn't have the heat to it that it used to. It's a mental kind of anger."* (p. 151) [30]. This patient's report appears to support the James-Lange theory. This patient suffered from a lesion, which influenced his patient biosignals and, simultaneously, his emotions have faded or are absent. However, the patient

still reported emotions, although differently. If biosignals determine emotions (completely); then, how can this be explained? Can such effects be assigned solely to higher level cognition (i.e., reasoning)? If not, this can be considered as support for the first notion of the Cannon-Bard theory. More than anything else this case once more illustrates the complexity of affective processes as well as the need for user identification, in particular research on special cases.

The second notion of the Cannon-Bard theory touches upon the core of *affective computing*. It suggests that the hunt of *affective computing* is one without a future. Cannon-Bard can be interpreted as that *affective computing* is of little use since biosignals do not show a singular mapping on specific emotions. On the one hand, nowadays, this statement is depicted as crude [25]. On the other hand, awareness emerges on that *affective computing* is very hard to bring to practice successfully [9].

Modern science has indeed confirmed that the number of sensory nerves indeed differs in distinct structures in human bodies (Cannon's notion 3). This can explain their internal variance to emotional sensitivity. Moreover, research has shown that cross-cultural and ethnic differences underly differences in people's biosignals. This was already shown by Sternbach and Tursky [65, 69] and, more recently, confirmed [53, 60, 63].

The fourth notion concerns the latency period of biosignals, which Cannon denoted as being 'long'. Indeed a response time is present with biosignals, which one could denote as being long. Moreover, it varies considerably between the several biosignals used with *affective computing* [1, 70]. The former is a problem, although in most cases a work around is, to some extent, possible. The latter is possibly even more important to take into account, when conducting *affective computing*. Regrettably, this is seldom done.

The fifth notion of Cannon stretches beyond biosignals. It concerns emotion's neurochemical dimension. This aspect of human physiology has a significant influence on experienced emotions. However, this dimension is not yet embraced by *affective computing* and, hence, falls beyond the scope of this article.

In 1962, Schachter and Singer [58] were the first to suggest that neither the theory of James-Lange nor the theory of Cannon-Bard was complete but both contained valuable elements. Nowadays, this is general opinion among neuroscientists [25]. More than anything else, these samples from history illustrate the knowledge already available as well as its lagoons, which should be embraced both by *affective computing*.

3 On the Complexity of Affect

Affect is inherently complicated and not well understood by science and engineering. In this section, I will illustrate this by approaching affect from three distinct angles. Starting with philosophy and, subsequently, psychology, I will go to AI and Human-Computer Interaction (HCI) and, finally, to computer science's I/O in relation to affect.

3.1 The Relation between Body and Mind[2]

As the old Greek already noted that *"the body could not be cured without the mind"*. Both are indisputably related and, thus, in principle, measurement of emotions should be feasible. Currently, this perspective gains acceptance and more and more relations between body and mind are unveiled. Recent work confirmed this relation. For example, when chronic stress is experienced, similar physiological responses emerge as were present during the stressful events from which the stress originates. If such physiological responses persist, they can cause pervasive and structural chemical imbalances in people's physiological systems, including their autonomic and central nervous system, their neuroendocrine system, their immune system, and even in their brain [17].

Although the previous enumeration of people's physiological systems can give the impression that we are close to a holistic model, it should be noted that this is in sharp contrast with the current level of science. For example, with (chronic) stress, a thorough understanding is still missing. This can be explained by the complexity of human's physiological systems, the continuous interaction of all systems, and their integral dynamic nature. However, [17] considers emotions as if these can be isolated and attributed to bodily processes only. Moreover, in relation to computing entities, the interaction consists of much more than emotions; however, the same is true when no computing is involved at all.

3.2 Cognitive Processes

There is more than AI with cognitive behavioral systems, there is also the interaction with their users. So, a HCI perspective has to be taken as well. In the 90s of the previous century, Nass and colleagues [46, 52] touched upon a new level of HCI: a personal, intimate, and emotional level (cf. [12, 14, 16]). Together with the work of [49, 50] their work positioned affective processes firmly as an essential ingredient of HCI.

The importance of affect for HCI can be well explained by denoting its influence on three cognitive processes, which are important in HCI context:

1. **Attention:** Affective processes take hold on several aspects of our cognitive processing [24] and, hence, HCI [71]. One of the most prominent effects of affect lies in its ability to capture attention. Affective processes have a way of being completely absorbing. Functionally, they direct and focus our attention on those objects and situations that have been appraised as important to our needs and goals [71]. This attention-getting function can be used advantageously in HCI context [62].
2. **Memory:** It should be noted that affect also has implications for learning and memory [8, 11]. Events with an affective load are generally remembered better than events without such a load, with negative events being dominant

[2] This subsection is almost verbatim identical with part of an online commentary on [31]; see also http://www.interaction-design.org/encyclopedia/affective_computing.html#egon+l.+van+den+broek

over positive events [52]. Further, affect improves memory for core information, while undermining memory for background information [39].

3. **Decision making:** Affective processes influence our flexibility and efficiency of thinking and problem solving [39]. It has also been shown that affect can (heavily) influence judgment and decision making [6, 61]. Affective processes tend to bias thoughts by way of an affect-filter. Thoughts are directed to a affect-consistent position, which can also increase the risk of distractions.

This triplet of cognitive processes illustrates that a careful consideration of affect in HCI can be instrumental in creating interfaces that are both efficient and effective as well as enjoyable and satisfying [62]. Moreover, the experience of emotions alters our experience in general [5, 31].

3.3 Affective Computing's I/O: Impressions / Expressions

In the introduction, I already stated that emotions and computers have become entangled and, in time, will inevitably embrace each other. Computer science and practice employs *input/output (I/O)* operations to characterize its processes. This notion can also be fruitfully utilized for *affective computing*, as I will illustrate here.

Table 1 shows a matrix that provides a characterization of machinery using, what could be, standard *I/O*. Machinery without any *I/O* (i.e., –/–) at all is of no use. In contrast, machinery without either input (i.e., *I*) or output (i.e., *O*) are common practice. However, most of us will assume both input and output (i.e., *I/O*), at least to a certain extent, with most of our machinery. For example, take our standard (office) PC with its output (i.e., at least a video (the screen) and audio) and its input (i.e., at least a keyboard and a pointing device). Emerging branches of science and engineering such as AI, AmI, and *affective computing*, however, aim to redecorate this traditional landscape and provide intuitive *I/O* handling. In the case of *affective computing*, what does this imply?

Computer science's notion of *I/O* operations can also be utilized to divide *affective computing* into four categories. In terms of *affective computing*, the output (*O*) denotes the expression of affect (or emotions) and the input (*I*) denotes the perception, impression, or recognition of affect. This division is adapted from the four cases, as they were identified by Rosalind W. Picard's in her thought-paper, which presented her initial thinking on *affective computing* [49]. Entities without any affective *I/O* (i.e., –/–), such as traditional machinery, can be very useful in all situations where emotions hinder instead of help. Entities with only affective *O* could for example be avatars, consumer products (e.g., a sports car), toys for children, and our TV. However, such entities would not know what affective state its user is in and, hence, what affect to show as they would lack the affective *I* for it. So, as its name, emotion-aware systems, already gives away, a requirement for such systems is affective *I*.

Only affective *I* is possible. In such cases, the affective *I* alters other processes (e.g., scheduling breaks for pilots) and no affective *O* is given but another type of output closes the system. In case of affective *I/O*, the affective *O* can follow

		O	
		no	yes
I	no	–/–	I/–
	yes	–/O	I/O

Table 1. A description of the four categories of *affective computing* in terms of computer science's input/output (I/O) operations. In terms of *affective computing*, I/O denotes the expression (O) and the perception, impression, or recognition (I) of affect. This division is adapted from the four cases identified by Rosalind W. Picard [49].

the affective I immediately or with a (fixed or varying) delay. The affective O can also take various forms. Moreover, the person who provides the affective I is not necessarily the person who receives the affective O.

The theoretical framework concerning affective processes is a topic of continuous debate. Consequently, an accurate interpretation of affective I and, subsequently, an appropriate affective O is hard to establish. In particular in real-world settings, where several sources of noise will disturb the closed-loop, this will be a challenging endeavor. So, currently, it is best to apply simple and robust mechanisms to generate affective O (e.g., on reflex agent level [56]) or slightly more advanced. Moreover, it is not specific states of affect that need to be the target but rather the core affect of the user that needs to be smoothly (and unnoticeably) directed to a target core state [34]. The next section will elaborate on core affect and related concepts.

4 In Search for Definitions

In 1993, Robert C. Solomon noted in the *Handbook of Emotions* (Chapter 1, p. 3, 1st ed.) [39] that " *"What is an emotion?"* is the question that *"was asked in precisely that form by William James, as the title of an essay he wrote for* Mind *well over* 100 *years ago (James,* 1884*). . . . But the question* "What is an emotion?" *has proved to be as difficult to resolve as the emotions have been to master. Just when it seems that an adequate definition is in place, some new theory rears its unwelcome head and challenges our understanding."* Regrettably, there is no reason to assume that this could not be the case for the concise theoretical framework that will be presented here (cf. [32]). Nevertheless, we need such a framework to bring emotion theory to *affective computing* practice.

4.1 Affect

In 2003, 10 years after Solomon's notion, in the journal *Psychological Review*, James A. Russell characterized the state-of-the-art of emotion (related) research as follows: *"Most major topics in psychology and every major problem faced by humanity involve emotion. Perhaps the same could be said of cognition. Yet, in the psychology of human beings, with passions as well as reasons, with feelings as well as thoughts, it is the emotional side that remains the more mysterious. Psychology and humanity can progress without considering emotion – about as fast as someone running on one leg."*)(p. 145) [55]. Where Solomon [39] (Chapter 1, p. 3, 1st ed.) stressed the complexity of affect and emotions, Russell [55] (p. 145) stressed the importance to take them into account. Indeed, affect and

emotions are of importance for psychology and humanity but *also* for (some branches of) science and engineering, as we will argue in this article.

Solomon's [39] (Chapter 1, p. 3, 1st ed.) and Russell's [55] (p. 145) quotes perfectly points towards the complexity of the constructs at hand (i.e., affect and emotion, amongst other things). It is well beyond the scope of this article to provide an exhaustive overview of theory on affect, emotion, and related constructs. However, a basic understanding and stipulative definitions are needed, as they are the target state *affective computing* is aiming at. This section will provide the required definitions. Since this article aims at *affective computing*, I will focus on affect as the key construct, which is, from a taxonomic perspective, a convenient choice as well. Affect is an umbrella construct that, instead of emotions, incorporates all processes I am interested in, as we will see in the remaining section.

Core affect is a neurophysiological state that is consciously accessible as a primitive, universal, simple (i.e., irreducible on the mental plane), nonreflective feeling evident in moods and emotions [51, 55]. It can exist with or without being labeled, interpreted, or attributed to any cause [55]. People are always and continuously in a state of core affect, although it is suggested that it disappears altogether from consciousness when it is neutral and stable [55]. Affect influences our attitudes, emotions, and moods and as such our feelings, cognitive functioning, behavior, and physiology [29, 55]; see also Table 2. As such, affect is an umbrella construct, a superordinate category [29].

Affect is similar to Thayer's activation [67], Watson and Tellegen's affect [72], and Morris' mood [45] as well as what is often denoted as a feeling [55]. As such, core affect is an integral blend of hedonic (pleasure-displeasure) and arousal (sleepy-activated) values; in other words, it can be conveniently mapped onto the valence-arousal model [38, 54, 55, 67]. However, note that the term "affect" is used throughout the literature in many different ways [51]. Often it is either ill defined or not defined at all. However, affect has also been positioned on another level than that just sketched; for example, as referring to behavioral aspects of emotion [29].

With affect being defined, we are left with a variety of related constructs. To achieve a concise but proper introduction to these constructs, we adopt Scherer's table of psychological constructs related to affective phenomena [10]; see Table 2. It provides concise definitions, examples, and seven dimensions on which the constructs can be characterized. Together this provides more than rules of thumb, it demarcates the constructs up to a reasonable and workable level. Suitable usage of Table 2 and the theoretical frameworks it relies on opens affect's black box and makes it a gray box [35, 48], which should be conceived as a huge progress.

4.2 Affective Computing

In 1995, Rosalind W. Picard wrote a technical report , which was a thought-paper that presented her initial thinking on *affective computing*. In a nutshell, this report identifies a number of crucial notions, which are still relevant. Moreover, Picard provided an initial definition of *affective computing: "... a set of ideas*

Table 2. Design feature delimitation of psychological constructs related to affective phenomena, including their brief definitions, and some examples. This table is adopted from [10].

construct	brief definition and *examples*	intensity	duration	synchro-nization	event focus	appraisal elicitation	rapidity of change	behavioral impact
Emotion	Relatively brief episode of synchronized response of all or most organismic subsystems in response to the evaluation of an external or internal event as being of major significance *(e.g., angry, sad, joyful, fewful, ashamed, proud, elated. desperate).*	++ → +++	+	+++	+++	+++	+++	+++
Mood	Diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause *(e.g., cheerful, gloomy, irritable, listless, depressed, buoyant).*	+ → ++	++	+	+	+	++	+
Inter-personal stances	Affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation *(e.g., distant, cold, warm, supportive, contemptuous).*	+ → ++	+ → ++	+	++	+	+++	++
Attitude	Relatively enduring, affectively colored beliefs, preferences, and predispositions towards objects or persons *(e.g., liking, loving, hating, valuing, desiring).*	0 → ++	++ → +++	0	0	+	0 → +	+
Personality traits	Emotionally laden, stable personality dispositions and behavior tendencies, typical for a person *(e.g., nervous, anxious, reckless, morose, hostile, envious, jealous).*	0 → +	+++	0	0	0	0	+

on what I call "affective computing," computing that relates to, arises from, or influences emotions." (p. 1) [49].

10 years later, Tao and Tan wrote a review on *affective computing* in which they defined it as: *"Affective computing is trying to assign computers the human-like capabilities of observation, interpretation and generation of affect features."* (cf. [66]). As such, they assured a one-on-one mapping of affect onto the traditional computer science / HCI triplet input (i.e., observation), processing (i.e., interpretation), and output (i.e., generation).

5 years later, Rafael A. Calvo and Sidney D'Mello [19] characterized the rationale of *affective computing* with: *"automatically recognizing and responding to a user's affective states during interactions with a computer can enhance the quality of the interaction, thereby making a computer interface more usable, enjoyable, and effective."*

I pose to adopt yet another definition, namely: *Affective computing is the scientific understanding and computation of the mechanisms underlying affect and their embodiment in machines.* This definition is inspired by the short definition of AI provided by the Association for the Advancement of Artificial Intelligence[3].

5 Discussion

Affective computing includes signal processing and machine learning techniques, which have a thorough mathematical foundation. Consequently, these elements are not those that slow down the progress of *affective computing*. In this article, I posed that *affective computing*'s true complexity lies in i) the definition of constructs related emotion, ii) their operationalization, and, subsequently, iii) their mapping on the signals available (e.g., audio, vision, or biosignals). To accelerate progress on *affective computing*, it is worth to become acquainted with the vast amount of work conducted in its related disciplines. Therefore, this article briefly touched on the history of *affective computing* (Section 2) and discussed the complexity of affect (Section 3). Subsequently, in Section 4 a concise set of definitions was provided on *affective computing*'s key concepts. In this section, I will identify lessons that can be learned from this endeavor and discuss some of challenges and limitations *affective computing* is facing.

The historical reflection in Section 2 was founded on the classical debate between the James-Lange theory and work of Cannon and Bard [25]. It is fair to pose that this debate is somewhat outdated and substantial progress has been made since the work of Cannon and Bard. Nevertheless, this debate still touches upon some key notions and identifies both knowledge and the gaps therein. Moreover, it provides a compact sketch of the true complexity underlying *affective computing*. As such, I hope that it can serve as an eye-opener and encourage readers to study the foundations of affective sciences in depth.

In Section 3, *affective computing* was approached from three independent angles: i) the relation between body and mind, ii) cognitive processes, and iii)

[3] Association for the Advancement of Artificial Intelligence (AAAI)'s URL: http://www.aaai.org/ [Last accessed on April 14, 2012].

affective computing's I/O. The first angle can be perceived as an extension of the debate presented in Section 2. It takes a philosophical stance, starting with the old Greek and ending with state of the art neuroscience research [17]. In a nutshell, I conclude that, despite the tremendous progress made in science and engineering, we are far from a complete understanding of the relation between our body, mind, and context [13, 15]. The second angle concerns the awareness that emotions can not be studied in isolation, they interact with (other) cognitive processes. However, as with the body and mind issue, this interaction is complex and our knowledge on this is thin. The third angle takes a computing perspective instead of an philosophical or cognitive perspective. *Affective computing* is described in terms of computer science's I/O, which was adopted from Rosalind W. Picard [49]. Taking this perspective helps in making explicit what we denote as *affective computing*, which is often a source of confusion.

Section 4 is dedicated to providing definitions of affect-related constructs and affective computing. It includes Table 2 that provides definitions and examples of emotion, mood, interpersonal stances, attitude, and personality traits and characterizes these constructs on seven dimensions. Although it is advisable to start research with well grounded definitions, a contrasting position can be taken as well. Already in 1941, Elizabeth Duffy published her article *"An explanation of 'emotional' phenomena without the use of the concept 'emotion' "* in which she starts by stating that she considers *"... 'emotion', as a scientific concept, is worse than useless. ... 'Emotion' apparently did not represent a separate and distinguishable condition."* (p. 283) [27]. Similar concerns were expressed in 1990, John T. Cacioppo and Louis G. Tassinary [18]. So, although this statement is 60 years old it is still (or, again) up to date, perhaps even more than ever (cf. [37]).

To ensure sufficient advancement, it has also been proposed to develop computing entities that respond on their user(s) physiological response(s), *without* the use of any interpretation of them in terms of emotions or cognitive processes [68]. This approach has been shown to be feasible for several areas of application (e.g., [34]). It suggests that emotion research has to mature further before affective computing can be brought to practice. This is a crude but honest conclusion for the field of affective computing. It implies that affective computing should take a few steps back before making its leap forward. Gross [29] summarized in his article *"The future's so bright, I gotta wear shades"* a list of hot topics on emotion research, which would be a good starting point for this process (see also [70]). I pose that if anything, affective computing has to learn more about its roots; then, affective computing can and probably will have a bright future!

Acknowledgements. I would like to thank Frans van der Sluis (Human Media Interaction (HMI), University of Twente, NL) and Joris H. Janssen (Philips Research, NL) for their comments on parts of this article. I thank the three anonymous reviewers for their valuable comments and constructive feedback. Further, I thank Lynn Packwood (HMI, UT, NL) for her careful proofreading. Finally, I would like to thank Anna Esposito (Second University of Naples and International Institute for Advanced Scientific Studies (IIASS), Italy) for asking

me to contribute to her volume. This publication was supported by the Dutch national program COMMIT (projects P4 Virtual worlds for well-being and P7 SWELL).

References

[1] van den Broek, E.L., et al.: Prerequisites for Affective Signal Processing (ASP) – Parts I–V. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BioSTEC 2009/2010/2011: Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies, Setubal, Portugal. INSTICC, Porto (2009/2010/2011)
[2] Backs, R.W., Boucsein, W.: Engineering Psychophysiology: Issues and applications. Lawrence Erlbaum Associates, Inc., Mahwah (2000)
[3] Bard, P.: On emotional expression after decortication with some remarks on certain theoretical views, Part I. Psychological Review 41(4), 309–329 (1934)
[4] Bard, P.: On emotional expression after decortication with some remarks on certain theoretical views, Part II. Psychological Review 41(5), 424–449 (1934)
[5] Barrett, L.F., Mesquita, B., Ochsner, K.N., Gross, J.J.: The experience of emotion. The Annual Review of Psychology 58, 373–403 (2007)
[6] Bechara, A.: The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. Brain and Cognition 55(1), 30–40 (2004)
[7] l'Abbé Bertholon, M.: De l'Électricité du corps humain. Tome Premiere, Lyon (1780)
[8] Blaney, P.H.: Affect and memory: A review. Psychological Bulletin 99(2), 229–246 (1986)
[9] Boehner, K., DePaula, R., Dourish, P., Sengers, P.: How emotion is made and measured. International Journal of Human-Computer Studies 65(4), 275–291 (2007)
[10] Borod, J.C.: The neuropsychology of emotion. Series in Affective Science. Oxford University Press, Inc., New York (2000)
[11] Bower, G.H., Gilligan, S.G., Monteiro, K.P.: Selectivity of learning caused by affective states. Journal of Experimental Psychology: General 110(4), 451–473 (1981)
[12] van den Broek, E.L.: Robot nannies: Future or fiction? nteraction Studies 11(2), 274–282 (2010)
[13] van den Broek, E.L.: Ubiquitous emotion-aware computing. Personal and Ubiquitous Computing 16 (in press, 2012)
[14] van den Broek, E.L., Lisý, V., Janssen, J.H., Westerink, J.H.D.M., Schut, M.H., Tuinenbreijer, K.: Affective Man-Machine Interface: Unveiling Human Emotions through Biosignals. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2009. CCIS, vol. 52, pp. 21–47. Springer, Heidelberg (2010)
[15] van den Broek, E.L., Schut, M.H., Westerink, J.H.D.M., Tuinenbreijer, K.: Unobtrusive Sensing of Emotions (USE). Journal of Ambient Intelligence and Smart Environments 1(3), 287–299 (2009)
[16] van den Broek, E.L., Westerink, J.H.D.M.: Biofeedback systems for stress reduction: Towards a bright future for a revitalized field. In: Proceedings of HealthInf 2012: International Conference on Health Informatics, Vilamoura, Algarve, Portugal, February 01-04, pp. 499–504. SciTePress, Portugal (2012)
[17] Brosschot, J.F.: Markers of chronic stress: Prolonged physiological activation and (un)conscious perseverative cognition. Neuroscience & Biobehavioral Reviews 35(1), 46–50 (2010)

[18] Cacioppo, J.T., Tassinary, L.G.: Inferring psychological significance from physiological signals. American Psychologist 45(1), 16–28 (1990)
[19] Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing 1(1), 18–37 (2010)
[20] Campbell, M., Joseph Hoane Jr., A., Hsu, F.: Deep Blue. Artificial Intelligence 134(1–2), 57–83 (2002)
[21] Cannon, W.B.: Bodily changes in pain, hunger, fear and rage: An account of recent researches into the function of emotional excitement. D. Appleton and Company, New York (1915)
[22] Cannon, W.B.: The James-Lange theory of emotion: A critical examination and an alternative theory. American Journal of Psychology 39(3-4), 106–124 (1927)
[23] Cochrane, P.: A measure of machine intelligence. Proceedings of the IEEE 98(9), 1543–1545 (2010)
[24] Compton, R.J.: The interface between emotion and attention: A review of evidence from psychology and neuroscience. Behavioral and Cognitive Neuroscience Reviews 2(2), 115–129 (2003)
[25] Dalgleish, T., Dunn, B.D., Mobbs, D.: Affective neuroscience: Past, present, and future. Emotion Review 1(4), 355–368 (2009)
[26] Davidson, R.J., Scherer, K.R., Hill Goldsmith, H.: Handbook of affective sciences. Oxford University Press, New York (2003)
[27] Duffy, E.: The conceptual categories of psychology: A suggestion for revision. Psychological Review 48(3), 177–203 (1941)
[28] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An overview of the DeepQA project. AI Magazine 31(3), 59–79 (2010)
[29] Gross, J.J.: The future's so bright, I gotta wear shades. Emotion Review 2(3), 212–216 (2010)
[30] Hohnmann, G.W.: Some effects of spinal cord lesions on experienced emotional feelings. Psychophysiology 3(2), 143–156 (1966)
[31] Höök, K.: Affective Computing: Affective Computing, Affective Interaction and Technology as Experience, ch. 12. Aarhus C., Denmark: The Interaction-Design.org Foundation (2012)
[32] Izard, C.E., et al.: Special section: On defining emotion. Emotion Review 2(4), 363–385 (2010)
[33] James, W.: What is an emotion? Mind 9(34), 188–205 (1884)
[34] Janssen, J.H., van den Broek, E.L., Westerink, J.H.D.M.: Tune in to your emotions: A robust personalized affective music player. User Modeling and User-Adapted Interaction 22(3), 255–279 (2012)
[35] Ju-Long, D.: Control problems of grey systems. Systems & Control Letters 1(5), 288–294 (1982)
[36] Kelley, T.D., Long, L.N.: Deep Blue cannot play checkers: The need for generalized intelligence for mobile robots. Journal of Robotics 2010, ID 5237571–8 (2010)
[37] Kleinginna, P.R., Kleinginna, A.M.: A categorized list of emotion definitions, with a suggestion for a consensual definition. Motivation and Emotion 5(4), 345–379 (1981)
[38] Lang, P.J.: The emotion probe: Studies of motivation and attention. American Psychologist 50(5), 372–385 (1995)
[39] Lewis, M., Haviland-Jones, J.M., Barrett, L.F.: Handbook of Emotions, 3rd edn. The Guilford Press, New York (2008)

[40] Lungarella, M., Iida, F., Bongard, J.C., Pfeifer, R. (eds.): 50 Years of Aritficial Intelligence. LNCS (LNAI), vol. 4850. Springer, Heidelberg (2007)
[41] Martin, I., Venables, P.H.: Techniques in Psychophysiology. John Wiley & Sons, Chichester (1980)
[42] Medawar, P.B.: Introduction and intuition in scientific thought, Memoir (Jayne lectures; 1968), vol. 075. Methuen & Co. Ltd., American Philosophical Society, London, Philidelphia (1969)
[43] Minsky, M.: The Society of Mind. Simon & Schuster Paperbacks, New York (1985)
[44] Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon & Schuster, New York (2006)
[45] Morris, W.N.: Mood: The frame of mind. Springer Series in Social Psychology. Springer, New York (1989)
[46] Nass, C.I., Moon, Y., Fogg, B.J., Reeves, B., Dryer, D.C.F.: Can computer personalities be human personalities? International Journal of Human-Computer Studies 43(2), 223–239 (1995)
[47] Neisser, U.: The imitation of man by machine – The view that machines will think as man does reveals misunderstanding of the nature of human thought. Science 139(3551), 193–197 (1963)
[48] Phan, L., Barriga, S.: Nico Frijda: Opening the Pandora's box of sciences. The Free Mind 1(2), 22–29 (2006)
[49] Picard, R.W.: Affective computing. Technical Report 321, M.I.T. Media Laboratory Perceptual Computing Section, Cambridge, MA, USA (1995)
[50] Picard, R.W.: Affective Computing. MIT Press, Boston (1997)
[51] Power, M.J., Dalgleish, T.: Cognition and emotion: From order to disorder. Hove, 2nd edn. Psychology Press/Taylor& Francis Group, East Sussex (2008)
[52] Reeves, B., Nass, C.: The media equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, New York (1996)
[53] Roberts, N.A., Levenson, R.W., Gross, J.J.: Cardiovascular costs of emotion suppression cross ethnic lines. International Journal of Psychophysiology 70(1), 82–87 (2008)
[54] Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)
[55] Russell, J.A.: Core affect and the psychological construction of emotion. Psychological Review 110(1), 145–172 (2003)
[56] Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall series in Artificial Intelligence, 3rd edn. Pearson Education, Inc., Upper Saddle River (2010)
[57] Sander, D., Scherer, K.R.: The Oxford companion to emotion and affective sciences. Series in Affective Science, 1st edn. Oxford University Press Inc., Oxford (2009)
[58] Schachter, S., Singer, J.E.: Cognitive, social, and physiological determinants of emotional state. Psychological Review 69(5), 379–399 (1962)
[59] Scherer, K.R., Bänziger, T., Roesch, E.B.: Blueprint for Affective Computing: A sourcebook. Series in Affective Science. Oxford University Press, Inc., New York (2010)
[60] Schneiderman, N., Weiss, S.M., Kaufmann, P.G.: Handbook of research methods in cardiovasular behavioral medicine. The Plenum series in Behavioral Psychophysiology and Medicine. Plenum Press, New York (1989)
[61] Schwarz, N.: Emotion, cognition, and decision making. Cognition & Emotion 14(4), 433–440 (2000)

[62] Sears, A., Jacko, J.A.: The Human-Computer Interaction handbook: Fundamentals, evolving technologies and emerging applications. In: Human Factors and Ergonomics, 2nd edn. Lawrence Erlbaum Associates/Taylor & Francis Group, LLC, New York (2008)
[63] Shen, B.J., Stroud, L.R., Niaura, R.: Ethnic differences in cardiovascular responses to laboratory stress: A comparison between Asian and white Americans. International Journal of Behavioral Medicine 11(3), 181–186 (2004)
[64] Simon, H.A.: Motivational and emotional controls of cognition. Psychological Review 74(1), 29–39 (1967)
[65] Sternbach, R.A., Tursky, B.: Ethnic differences among housewives in psychophysical and skin potential responses to electric shock. Psychophysiology 1(3), 241–246 (1965)
[66] Tao, J., Tan, T.: Affective Information Processing. Springer, London (2009)
[67] Thayer, R.E.: The biopsychology of mood and arousal. Oxford University Press, Inc., New York (1989)
[68] Tractinsky, N.: Tools over solutions? Comments on interacting with computers special issue on affective computing. Interacting with Computers 16(4), 751–757 (2004)
[69] Tursky, B., Sternbach, R.A.: Further physiological correlates of ethnic differences in responses to shock. Psychophysiology 4(1), 67–74 (1967)
[70] van den Broek, E.L.: Affective Signal Processing (ASP): Unraveling the mystery of emotions. Ph.D. thesis, Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede, The Netherlands (2011)
[71] Vertegaal, R.: Attentive user interfaces. Communications of the ACM 46(3), 30–33 (2003)
[72] Watson, D., Tellegen, A.: Toward a consensual structure of mood. Psychological Bulletin 98(2), 219–235 (1985)

Author Index