

IFIP AICT 425



Yves Rybarczyk  
Tiago Cardoso  
João Rosas

Luis M. Camarinha-Matos  
(Eds.)

# Innovative and Creative Developments in Multimodal Interaction Systems

9th IFIP WG 5.5 International Summer Workshop  
on Multimodal Interfaces, eNTERFACE 2013  
Lisbon, Portugal, July 15 – August 9, 2013  
Proceedings

 Springer

Editor-in-Chief

*A. Joe Turner, Seneca, SC, USA*

Editorial Board

Foundations of Computer Science

*Jacques Sakarovitch, Télécom ParisTech, France*

Software: Theory and Practice

*Michael Goedicke, University of Duisburg-Essen, Germany*

Education

*Arthur Tatnall, Victoria University, Melbourne, Australia*

Information Technology Applications

*Erich J. Neuhold, University of Vienna, Austria*

Communication Systems

*Aiko Pras, University of Twente, Enschede, The Netherlands*

System Modeling and Optimization

*Fredi Tröltzsch, TU Berlin, Germany*

Information Systems

*Jan Pries-Heje, Roskilde University, Denmark*

ICT and Society

*Diane Whitehouse, The Castlegate Consultancy, Malton, UK*

Computer Systems Technology

*Ricardo Reis, Federal University of Rio Grande do Sul, Porto Alegre, Brazil*

Security and Privacy Protection in Information Processing Systems

*Yuko Murayama, Iwate Prefectural University, Japan*

Artificial Intelligence

*Tharam Dillon, Curtin University, Bentley, Australia*

Human-Computer Interaction

*Jan Gulliksen, KTH Royal Institute of Technology, Stockholm, Sweden*

Entertainment Computing

*Matthias Rauterberg, Eindhoven University of Technology, The Netherlands*

## **IFIP – The International Federation for Information Processing**

IFIP was founded in 1960 under the auspices of UNESCO, following the First World Computer Congress held in Paris the previous year. An umbrella organization for societies working in information processing, IFIP's aim is two-fold: to support information processing within its member countries and to encourage technology transfer to developing nations. As its mission statement clearly states,

IFIP's mission is to be the leading, truly international, apolitical organization which encourages and assists in the development, exploitation and application of information technology for the benefit of all people.

IFIP is a non-profitmaking organization, run almost solely by 2500 volunteers. It operates through a number of technical committees, which organize events and publications. IFIP's events range from an international congress to local seminars, but the most important are:

- The IFIP World Computer Congress, held every second year;
- Open conferences;
- Working conferences.

The flagship event is the IFIP World Computer Congress, at which both invited and contributed papers are presented. Contributed papers are rigorously refereed and the rejection rate is high.

As with the Congress, participation in the open conferences is open to all and papers may be invited or submitted. Again, submitted papers are stringently refereed.

The working conferences are structured differently. They are usually run by a working group and attendance is small and by invitation only. Their purpose is to create an atmosphere conducive to innovation and development. Refereeing is also rigorous and papers are subjected to extensive group discussion.

Publications arising from IFIP events vary. The papers presented at the IFIP World Computer Congress and at open conferences are published as conference proceedings, while the results of the working conferences are often published as collections of selected and edited papers.

Any national society whose primary activity is about information processing may apply to become a full member of IFIP, although full membership is restricted to one society per country. Full members are entitled to vote at the annual General Assembly, National societies preferring a less committed involvement may apply for associate or corresponding membership. Associate members enjoy the same benefits as full members, but without voting rights. Corresponding members are not represented in IFIP bodies. Affiliated membership is open to non-national societies, and individual and honorary membership schemes are also offered.

Yves Rybarczyk Tiago Cardoso  
João Rosas Luis M. Camarinha-Matos (Eds.)

# Innovative and Creative Developments in Multimodal Interaction Systems

9th IFIP WG 5.5 International Summer Workshop  
on Multimodal Interfaces, eNTERFACE 2013  
Lisbon, Portugal, July 15 – August 9, 2013  
Proceedings



Springer

## Volume Editors

Yves Rybarczyk

Tiago Cardoso

João Rosas

Luis M. Camarinha-Matos

Universidade Nova de Lisboa

Departamento de Engenharia Electrotécnica

Quinta da Torre, 2829-516 Monte de Caparica, Portugal

E-mail: {yr, tomfc, jrosas, cam}@uninova.pt

ISSN 1868-4238

ISBN 978-3-642-55142-0

DOI 10.1007/978-3-642-55143-7

e-ISSN 1868-422X

e-ISBN 978-3-642-55143-7

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014936102

© IFIP International Federation for Information Processing 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

*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)

# Preface

This book contains the proceedings of the 9th International Summer Workshop on Multimodal Interfaces (eNTERFACE 2013), which was held in Lisbon, Portugal, during July 15th to August 9th, 2013. Following the tremendous success of the previous eNTERFACE workshops ([www.enterface.net](http://www.enterface.net)) held in Mons, Belgium (2005), Dubrovnik, Croatia (2006), Istanbul, Turkey (2007), Paris, France (2008), Genoa, Italy (2009), Amsterdam, The Netherlands (2010), Plzen, Czech Republic (2011) and Metz, France (2012), eNTERFACE 2013 aimed at continuing and enhancing the tradition of collaborative, localized research and development work by gathering, in a single place, leading researchers in multimodal interfaces and students to work on specific projects for four complete weeks. In this respect, it is an innovative and intensive collaboration scheme, designed to allow researchers to integrate their software tools, deploy demonstrators, collect novel databases, and work side by side with a great number of experts. It is thus radically different from traditional scientific workshops, in which only specialists meet for a few days to discuss state-of-the-art problems, but do not really work together.

In 2013, more than seventy researchers participated in eNTERFACE, which confirmed it as the largest workshop on multimodal interfaces. The event was attended by senior researchers, who were mainly university professors, industrial or governmental researchers presently working in widely dispersed locations, and junior researchers, who were mostly PhD students. In the first phase of the workshop, a call for proposals was circulated, for which interested researchers submitted project ideas. This was an international call, and it was widely circulated in all related major scientific networks. The Scientific Committee evaluated the proposals, and selected nine projects. In the second phase, a call for participation was circulated, in which the project leaders got to build their team. The would-be participants sent in a CV and why they were interested in joining a project. A small number of under-graduates were also selected, set on outstanding academic promise. Graduate students familiar with the field were selected in accordance with their demonstrated performance. This year, we have targeted projects on Innovative and Creative Developments in Multimodal Interaction Systems.

All the eNTERFACE 2013 projects have tackled new trends in human-machine interaction (HMI). The way in which the individual interacts with the devices is changing, mainly because of the boom of the gaming industry. This change encompasses two aspects. First of all, the user interfaces are more and more natural; and the large number of projects using a Kinect is a good example of this fact. The current HMI does not involve a single kind of input/output anymore, but a sophisticate signal processing from a combination of sensorial modalities and motor skills. Second, the development of applications based on

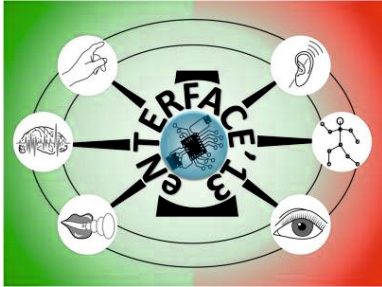
innovative scenario inspired from game concepts, for professional purposes (e.g. serious games), has significantly increased during the last decade. To address such a tendency, the studies presented in this volume have required a thin collaboration between computer scientists who have developed the prototypes, and other parties who have asked the relevant questions (e.g. artists, psychologists, industrial partners). The nine articles that compose this book are organized in two topical sections. The first one presents different proposals focused on some fundamental issues regarding multimodal interactions (i.e. telepresence, speech synthesis and interactive agents modeling). The second is a set of development examples in key areas of HMI applications (i.e. education, entertainment, and assistive technologies).

We would like to thank the authors for their contribution and the Steering Committee of eNTERFACE for the reviewing process of the articles, especially Albert Ali Salah, Gualtiero Volpe, Igor Pandzic, Antonio Camurri and Olivier Pietquin. We are also grateful to the sponsors of the event: Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, UNINOVA, EUCog network and IFIP. Finally, we also wish to express our thanks to Springer for making the publication of this book possible.

January 2014

Yves Rybarczyk  
Tiago Cardoso  
João Rosas  
Luis M. Camarinha-Matos

# Organization



## 9<sup>th</sup> International Summer Workshop on Multimodal Interfaces

Lisbon, Portugal,  
15 July – 9 August 2013

### Workshop Chair

Yves Rybarczyk

Portugal

### Organizing Committee Co-Chairs

Tiago Cardoso

Portugal

João Rosas

Portugal

### International Program Committee

Albert Ali Salah

University of Amsterdam, The Netherlands

Antonio Camurri

University of Genova, Italy

Benoît Macq

Université Catholique de Louvain, Belgium

Bülent Sankur

Bogazici University, Turkey

Christophe d'Alessandro

CNRS-LIMSI, France

Gualtiero Volpe

University of Genova, Italy

György Kovács

MTA SZTAKI, Hungary

Igor Pandzic

Zagreb University, Croatia

Luis M. Camarinha-Matos

New University of Lisbon, Portugal

Miloš Železný

University of West Bohemia, Czech Republic

Olivier Pietquin

Metz Supélec, France

Peter Bertok

RMIT University, Australia

Thierry Dutoit

Faculté Polytechnique de Mons, Belgium

Yves Rybarczyk

New University of Lisbon, Portugal



## Technical Sponsors



*EUCog - European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics*



*IFIP WG 5.5 COVE - Cooperation Infrastructure for Virtual Enterprises and electronic business*

## Organizational Sponsors



## Organized by:

Department of Electrotechnical Engineering FCT/UNL

# Table of Contents

## Part I: Fundamental Issues

Body Ownership of Virtual Avatars: An Affordance Approach of Telepresence . . . . .	3
<i>Tiago Coelho, Rita de Oliveira, Tiago Cardoso, and Yves Rybarczyk</i>	
Reactive Statistical Mapping: Towards the Sketching of Performative Control with Data . . . . .	20
<i>Nicolas d’Alessandro, Joëlle Tilmanne, Maria Astrinaki, Thomas Hueber, Rasmus Dall, Thierry Ravet, Alexis Moinet, Huseyin Cakmak, Onur Babacan, Adela Barbulescu, Valentin Parfait, Victor Huguenin, Emine Sümeyye Kalaycı, and Qiong Hu</i>	
Laugh When You’re Winning . . . . .	50
<i>Maurizio Mancini, Laurent Ach, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, Debajyoti Datta, Yu Ding, Stéphane Dupont, Harry J. Griffin, Florian Lingenfelser, Radoslaw Niewiadomski, Catherine Pelachaud, Olivier Pietquin, Bilal Piot, Jérôme Urbain, Gualtiero Volpe, and Johannes Wagner</i>	
Tutoring Robots: Multiparty Multimodal Social Dialogue with an Embodied Tutor . . . . .	80
<i>Samer Al Moubayed, Jonas Beskow, Bajibabu Bollepalli, Ahmed Hussen-Abdelaziz, Martin Johansson, Maria Koutsombogera, José David Lopes, Jekaterina Novikova, Catharine Oertel, Gabriel Skantze, Kalin Stefanov, and Gül Varol</i>	
Touching Virtual Agents: Embodiment and Mind . . . . .	114
<i>Gijs Huisman, Merijn Bruijnes, Jan Kolkmeier, Merel Jung, Aduén Darriba Frederiks, and Yves Rybarczyk</i>	

## Part II: Key Applications

Kinect-Sign: Teaching Sign Language to “Listeners” through a Game . . .	141
<i>João Gameiro, Tiago Cardoso, and Yves Rybarczyk</i>	
Hang in There: A Novel Body-Centric Interactive Playground . . . . .	160
<i>Robby van Delden, Alejandro Moreno, Carlos Ramos, Gonçalo Carrasco, Dennis Reidsma, and Ronald Poppe</i>	

KINterestTV - Towards Non-invasive Measure of User Interest While Watching TV . . . . .	179
<i>Julien Leroy, François Rocca, Matei Mancias, Radhwan Ben Madhkour, Fabien Grisard, Tomas Kliegr, Jaroslav Kuchar, Jakub Vit, Ivan Pirner, and Petr Zimmermann</i>	
Development of an Ecosystem for Ambient Assisted Living . . . . .	200
<i>João Rosas, Luis M. Camarinha-Matos, Gonçalo Carvalho, Ana Inês Oliveira, and Filipa Ferrada</i>	
<b>Author Index . . . . .</b>	<b>229</b>

**Part I**  
**Fundamental Issues**

# Body Ownership of Virtual Avatars: An Affordance Approach of Telepresence

Tiago Coelho<sup>1</sup>, Rita de Oliveira<sup>2</sup>, Tiago Cardoso<sup>1</sup>, and Yves Rybarczyk<sup>1,\*</sup>

<sup>1</sup>Electrotechnical Engineering Department, New University of Lisbon, Portugal  
{yr, tcoelho, tomfc}@uninova.pt

<sup>2</sup>Department of Applied Sciences, London South Bank University, UK  
r.oliveira@lsbu.ac.uk

**Abstract.** Virtual environments are an increasing trend in today's society. In this scope, the avatar is the representation of the user in the virtual world. However, that relationship lacks empirical studies regarding the nature of the interaction between avatars and human beings. For that purpose it was studied how the avatar's modeled morphology and dynamics affect its control by the user. An experiment was conducted to measure telepresence and ownership on participants who used a Kinect Natural User Interface (NUI). The body ownership of different avatars was assessed through a behavioral parameter, based on the concept of affordances, and a questionnaire of presence. The results show that the feelings of telepresence and ownership seem to be greater when the kinematics and the avatar's proportions are closer to those of the user.

**Keywords:** Avatar, telepresence, ownership, affordances, natural user interface, virtual environment.

## 1 Introduction

### 1.1 Telepresence and Ownership

Telepresence is the feeling of being present in a place where the person is not [1]. This feeling can be achieved while an individual is using a simulator or performing a certain task in a virtual environment such as a game [2]. Another way this phenomenon can occur is in teleoperation, in which the user remote-controls a robot with a camera that provides the teleoperator with visual feedback from the working space [3]. Telepresence is a critical mental process as it increases the immersion of the individual within a certain task. Teleoperation and virtual environments are the most common situations in which a feeling of telepresence may occur.

One of the most specific cases of telepresence is body ownership, in which the individual is so immerse in the teleoperation task he/she is performing that he/she

---

\* Departamento de Engenharia Electrotécnica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre 2829-516 Monte de Caparica, Portugal  
email: yr@uninova.pt

believes the remote machine is part of him/her [4]. In an experiment carried out by Sumioka et al. [5], subjects remote-controlled a human-like machine. They had a first person view over the machine, which replicated every move of the individual. The participant's reactions were gauged by measurement of the skin conductance. The results show that the participants seemed to feel the machine as it was their own body. In addition, the feeling of ownership can also happen in other cases of mediated situations such as in virtual reality, in which the individual believes he/she is the avatar.

The study that revealed for the first time the phenomenon of ownership is the Rubber Hand Illusion [6]. In this experiment, the participant's hand is hidden and a rubber hand is visible in its place. A tactile stimulation is applied simultaneously to both hands. After a while, the individual has the feeling that the fake hand is his/her own. A similar effect has also been observed in virtual reality [7, 8]. Ehrsson et al. [9] recorded the brain activity of participants when they were submitted to the same experiment. The results showed a significant activation of the parietal cortex only in the presence of a synchronous and congruent visuo-tactile stimulation of the rubber and the real hand. In addition, a positive correlation between the physiological and ownership questionnaire data confirmed that the participants were considering the rubber hand as their own hand.

It seems that the ownership feeling is not exclusive of the individual limbs and can occur on the entire body. In Petkova and Ehrsson's [10] experiment, participants wore a head-mounted display and had a first person view over a body-sized mannequin. They received visual and tactile stimulations over the whole body. In that condition, the participants had the feeling that the mannequin's body was their own. The ownership feeling was measured through skin conductance, which can detect psychological or physiological alterations. Authors stress the fact that a human-like representation of the mannequin and a synchronous visuo-tactile stimulation were crucial to trigger the ownership illusion.

A more surprising observation is the fact that body ownership may also occur in a simple situation of tool-use. Studies showed that when individuals manipulate an artifact, they consider it as an extension of their arm [11]. The initial study was performed with non-human primates and their brain activity was measured by electrodes. The results show that some specific bimodal neurons coding for the monkey's hand fire in a similar manner when a stimulus is applied to the hand or close to the tool manipulated by the animal [12]. This is clear evidence showing that the artifact seems to be integrated into the primate body schema.

Overall, these experiments showed that ownership occurs in the brain, after integration of multimodal information (vision, touch and proprioception) in order to build a coherent representation of the body.

## 1.2 Methods for Measuring Telepresence

**Questionnaires.** There are several ways to measure telepresence. One way the evaluation can be done is by questionnaire, in which the user answers a few questions in order to express what they felt during the experiment. This is probably the most popular kind of presence assessment. A significant number of studies on telepresence or

ownership ask their participants to fill a questionnaire when the experiment is over [6, 9, 10, 13].

Questionnaires are mostly used because of the simplicity of their implementation and the large range of possible questions. They are also a very quick and practical way for people to express their feelings, on a numerical scale that allows quantification and comparisons. Nevertheless there are some disadvantages, such as misinterpretation of questions, subjectivity of answers, the scale level number (odd vs. even) or, since it happens after the experiment, participants might forget what they felt. Another disadvantage is the number of questions: if there are too many, the participants may lose interest and answer randomly.

**Physiological Parameters.** Another way to gauge presence is by physiological parameters such as heart rate, galvanic skin response, electromyography or electroencephalogram. The galvanic skin response (GSR) measures the skin conductance of electricity. The variation of skin conductance occurs by changes in the moisture of the skin. Emotional stimulus triggers the sympathetic nervous system to increase the activity of the sweat glands. Armel and Ramachandran [14] measured the skin conductance in their rubber hand illusion experiment. They threatened to harm the rubber hand. If the participant thought that the rubber hand was his/her own, the skin conductance results showed signs of arousal.

The electromyography (EMG) measures the electric activity of activated muscles. The signals can also be used to detect neurologic activity. Slater et al. [15] used this technique during a rubber hand illusion experiment carried out in virtual reality. When ownership was achieved, the virtual hand was twisted. The EMG showed the twist induced a motor activity along the participant's real arm.

The electroencephalogram (EEG) directly measures electric activity of the brain in a non-invasive way. González-Franco et al. [16] showed that EEG can be used to assess presence in a virtual environment. They conducted an experiment in which the arm of the avatar was threatened. Results showed that when facing a threat, the participants lowered their motor cortex activity.

**Behavioral Assessment: Affordances.** In this project, affordances are used as a behavioral assessment to measure telepresence in a virtual environment. Affordances are a concept first suggested in the literature by J.J. Gibson [17]. An affordance is an action possibility whereby people perceive their environment and the objects within it as possibilities of doing certain actions and not doing other actions. Affordances exist where the characteristics of the object and the characteristics of the person match in a particular way. For instance, most chairs will afford sitting to most adults, but will not afford sitting to a 6-months baby, and might afford standing to someone making a speech. This concept is also applicable to other animal species, like for example a tree can afford nourishment to a giraffe but for a bird it can afford nesting. An affordance is a combination of the physical characteristics of the object and the animal, the knowledge about the object, and the needs to the animal at a particular time. In some cases, the action possibility may be harmful, in which case the animal may choose not to perform the action. For example a knife affords cutting into various surfaces

because it has a blade. If someone grasps it by the handle it affords cutting into bread or cut through paper but it also affords injury if grasped by the blade. Another example is about apertures. An aperture will only afford passage if it is wider than the individual. If it is narrower it may afford passage if the individual rotates upon himself [18]. Affordances are based on experience, in the sense that people learn to perceive the relevant characteristics of the environment and objects within it. This means they will be common to many individuals (e.g., passing through apertures which are large enough) but different from one individual to another (e.g., a rugby player, a gymnast, or a child will fit through different apertures). The link between perception and action which guides people's decision evolves over time through experience [19].

After the initial study by Warren and Whang [18] testing affordances, other studies have followed which explore and test the notion of affordances, such as [20, 21]. One crucial finding was that the possibilities for action available to an individual are scaled to the individual's body. This scaling factor is important because it links object properties and individual's dimensions through an invariant value; this means there is a lawful relation underpinning (at least some) affordances. Such lawful relations have been found in various animals. In human participants, this was found in stair climbing where participants deem a stair climbable (without the aid of hands) if the raiser is smaller than 0.88 their leg length [22]. This was also found to be the case in passing through apertures where participants rotate their shoulders over their longitudinal axis if the aperture is smaller than 1.4 the width of their shoulders [18].

## 2 Questioning and Prototype Solution

### 2.1 Problematic

The main purpose of this article is to evaluate the effect of an avatar on the feelings of telepresence. Avatars are very common in virtual environments. They are the alter-egos of users in the virtual world. The avatar is usually seen from a first or third person perspective. Several aspects can be studied regarding the avatar, such as its dynamics, morphology or physical appearance. In addition, aspects such as the camera perspective or a visuo-motor feedback from hardware might change the ownership feelings about the avatar. The physical user interface is an important parameter to take into account because the control the user has on the avatar influences his/her behavior [23]. All of these features may influence the way a person feels towards the avatar [24]. Here, the avatar's *morphology* and *dynamics* were studied when a human user controlled it through a NUI.

The *morphology* of the avatar is an important feature in achieving body ownership. For instance, Tsakiris and Haggard [25] carried out the rubber hand illusion experiment with a fake hand and with a wooden stick. Their results showed that, with the rubber hand, ownership was easier to achieve than with the wooden stick. Because feelings of ownership happened with the hand, a similar result is expected if applied on a larger scale to a scaled-to-user avatar. This was done by Petkova and Ehrsson [10] in a study using a camera on a mannequin in the real world.



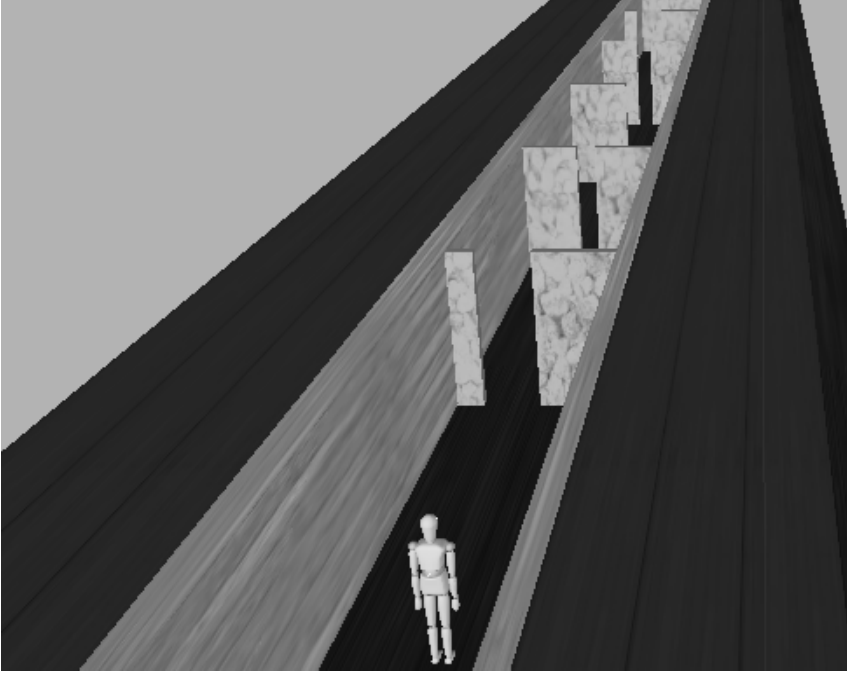
The *dynamics* of the avatar may also help to induce the feelings of telepresence and ownership on its user. If there is real-time congruence between the movements of the user and the movements of the avatar, then the feelings of telepresence and ownership should be greater than with incongruent movements. This is supported by the experiment carried out by Kalckert and Ehrsson [26]. These authors showed that the rubber hand illusion can be induced through a simple visuomotor correlation, without the need for a tactile stimulation as had been used in the original study of Botvinik and Cohen [6].

The study presented here is composed with two experimental conditions. In one of the conditions, the avatars are morphologically proportional (similar) to each participant. It is possible to have a dynamic avatar fully proportional to the user thanks to a full body motion capture. In the other condition, the avatar resembles the first one in how it looks, but it is the same (standard) for every participant. In addition, the movements of this standard avatar do not exactly match the participant's movements, as it only moves sideways and rotates upon itself. Aside from the avatar conditions, there was also a speed condition: fast and slow.

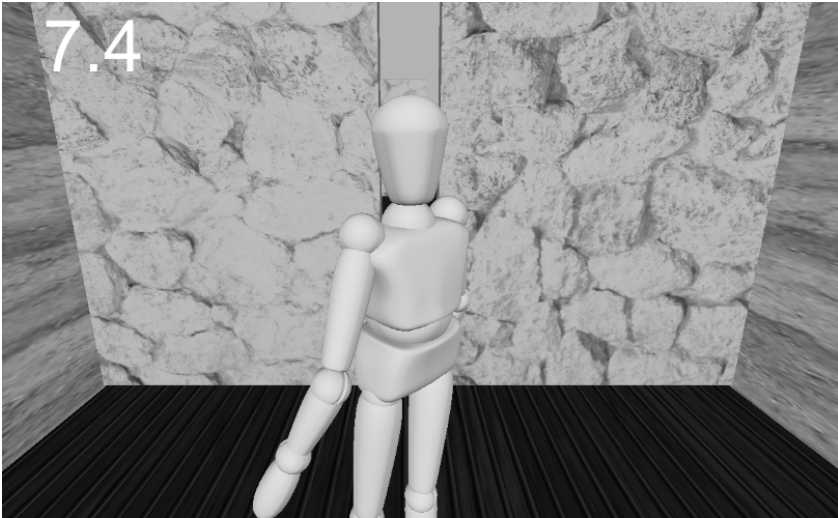
## 2.2 Developed Tool

The ATTAVE (Avatar Telepresence Testing: Affordances in Virtual Environments) is the name of the prototype developed. The ATTAVE is a virtual environment where the avatars exist and where experiments using affordances are performed. This virtual environment will be the same for both avatar conditions. Thus, the only parameters that change in the experiment are the avatar's characteristics. The design of the prototype is based on a study performed by Warren and Whang [18]. The authors evaluated how people passed through apertures considering the shoulder width of the participant and the ratio between the aperture and their shoulder width. The participants passed through several apertures of various sizes and the degree of their shoulder rotation was recorded. There were two speed conditions: a slow and a fast walking speed. Results showed that the participants only walked frontally through the aperture when the ratio between the aperture and the shoulders was smaller than 1.4. The present study will be similar to the one performed by Warren and Whang [18], but will be performed in a virtual environment. By replicating a study performed in the real world, telepresence can be verified if the same behaviors that happen in the real world also happen in the virtual environment.

The display of ATTAVE consists of a virtual scenario showing a long treadmill moving towards a visible avatar (and also towards the participant). The avatar resembles a wooden mannequin and is visible from head to knees, as the viewpoint of the participant is 2 m behind the avatar. The treadmill is enclosed on the side by tall walls. On the treadmill itself there are frontal green walls with an aperture on the left, centre or right side of the wall. All surfaces have texture as can be seen in Figure 1. The participants could control the translation and rotation of the avatars by physically moving side to side and rotating their own shoulders. Shoulder rotation proportionally slowed down the treadmill. The participants' task was to avoid collisions and pass through all doors as fast as possible.



**Fig. 1.** ATTAVE seen from above (this was not the perspective used in the study)



**Fig. 2.** Virtual avatar and environment viewed from the participants' perspective

The software chosen for modeling the avatars was Blender. The final models resembled a wooden mannequin (Figure 2). The software used to create the animation of the (standard) avatar was iPi Recorder and iPi Mocap Studio (iPi Soft). This tool was

chosen because it can record the movements captured by the Kinect and associate those movements to the avatar. With this method, an animation was created of a sidestep with a human model performing the action. This was the standard side-step used by the avatar in the standard condition.

Unity 3d was the game engine chosen for designing ATTAVE and running the application. The advantage of Unity 3d is its MonoDevelop IDE (Integrated Development Environment) integrated in the software, which facilitates the scripting part of the game. The scripting was made entirely in Java scripting language for Unity. Another reason that led to the use of Unity 3d in the project was its easy integration with the Kinect, which is made possible thanks to a framework called OpenNI (Open Source Natural Interaction) provided by Primesense.

For the full body motion capture a Kinect NUI was used. This device uses an optic technology that allows detection of the human body thanks to an infrared depth camera. This choice was due to its ease to set up, as it can be ready to use in less than five minutes, and also due to its low cost compared to other equipment of the same category. Another advantage in using this system is the fact that the user does not have to wear a specific suit and, consequently, he/she has full range of movement. In order to provide the participant with audio-visual feedback, a noise and a flash next to the avatar's shoulders were displayed whenever there was a collision with a wall.

When the participant rotated the shoulders, the speed of the treadmill decreased in proportion to the cosine of the angle between the shoulders' axis and the avatar's translation axis, as described by the formula below.

$$\text{current speed} = \text{set speed} * (1 - 0.4 * \sin(\text{angle})) \quad (1)$$

In (1), the value of the angle is taken as 0 if the individual is in a frontal position towards the door and 90° if the individual is facing the side walls. The angles are in absolute values from 0° to 90°. The speed decrease with rotation was introduced because it simulates a normal behavior observed in the real world. People usually reduce their locomotion velocity when they have to pass through a narrow aperture. The decreasing value of 0.4, used to calculate the current speed, was based on pilot trials. Moreover, the correlation between rotation and speed was an incentive for participants not to rotate their shoulders unless there was a danger of collision, as the rotation slowed down the treadmill and added to their total time on the task.

## 3 Experiment

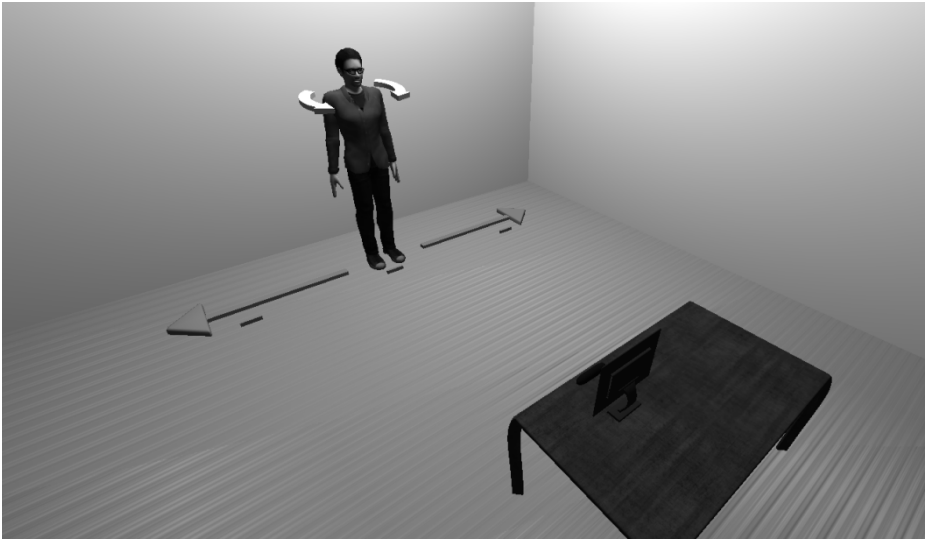
### 3.1 Participants

Participants were 24 university students (18 male and 6 female, aged between 20 and 28), with normal or corrected-to-normal vision and varied experience in playing video games. Half of the participants performed in two conditions (similar fast, similar slow) and the other half performed in two other conditions (standard fast, standard slow). This was done to enable the study of eventual learning effects under each speed

condition. The experiments were approved by the local ethics committee of the Nova University Lisbon.

### 3.2 Setup

The experiment was conducted in a  $3 \times 3$  m area. Participants stood 3 m away from a 75 cm height table. On the table was mounted an off-the-shelf Kinect sensor (Microsoft, for Xbox 360) and an 18" computer screen (1440  $\times$  900 pixel resolution), both connected to a PC. Three small marks on the floor indicated the positions aligned with the three apertures on the display (see Figure 3).



**Fig. 3.** Physical setup of the experiment

### 3.3 Experimental Design

Participants performed 32 trials for each of 2 speed conditions and for each of 4 sessions. Also, there were 2 avatar conditions; each used in a group of participants. The 32 trials consisted of apertures that showed in the central position with widths gradually increasing relative to the avatar's shoulders from 0.7 to 2.2 and then gradually decreasing from 2.2 to 0.7 (in steps of 0.1). When the avatar passed through each of these apertures, the value corresponding to the angle between the shoulders was recorded. These trials were alternated with 32 dummy trials with apertures of constant size shown in the right and left side positions. These side apertures remained twice the shoulder width of the avatar and were not used for data collection. Every aperture was 10 meters away from the next aperture. The two speed conditions were slow and fast (respectively 5 and 10 Km/h). They were taken from the walking speeds reported by Warren and Whang [18] and adjusted during pilot testing. The avatar condition consisted of manipulating the morphology and movements of the avatar. In the *similar*

avatar condition the avatar was anatomically proportional to the dimensions of the participant and all segments were animated to mimic the natural movements of the participant. In the *standard* condition the dimensions of the avatar were standard for all participants and the avatar had only two degrees of freedom: translation sideways and rotation upon itself (see arrows on Figure 3). An animation of a sidestep was implemented on the avatar when the participant performed a sidestep. This animation was recorded with a natural user interface and results from a sidestep performed by a human being.

The ratio between each virtual door and the avatar's shoulders width was the independent variable manipulated. The dependent variable was the angle between the shoulders during the passage through each aperture.

### 3.4 Procedure

The experiment started with participants reading and signing the consent form. Then, the Kinect was calibrated to the participants' movements. Participants were instructed to avoid collisions and complete the task in the shortest possible time, and were informed that shoulder rotation proportionally decreased the speed of the treadmill. In each session, participants completed the increasing-decreasing series in the slow condition followed by the fast condition. Participants were asked to fill in a questionnaire adapted from Witmer et al. [27]. Finally, the measure of participants' height and shoulder width was taken. In total, each session lasted about 20 minutes.

### 3.5 Results

**Data Analysis.** The main dependent variable was the critical ratio after which the participant passed without rotation. Following Warren and Whang [18], the two values of the critical ratio of shoulder rotation from the increasing-decreasing series were averaged. The critical ratio was that with an angle smaller than  $16^\circ$  and after which all angles were smaller than  $16^\circ$  (one exception was permitted provided the angle was smaller than  $40^\circ$  and the average angles remained smaller than  $16^\circ$ ). A critical ratio was calculated for each participant, condition, and session and these were used in the data analysis.

To examine learning effects from session to session, the critical ratios were submitted to a multivariate repeated measures analysis of variance (MANOVA) with the factor session (4 levels), and using the 4 conditions as measures (slow-similar, fast-similar, slow-standard, fast-standard). Based on the results of this analysis, the averages of the last 3 sessions were used in the remaining analysis.

To examine the effect of conditions on the critical ratios, the individual critical ratios from the last 3 sessions were averaged and submitted to a repeated measures analysis of variance (ANOVA) with factors speed (2 levels: slow and fast) and avatar (2 levels: similar and standard). The same analysis was conducted for the total durations, that is, the time that the participants took to complete the task by going through all the apertures. The same analysis was also used to test the effects on collisions.

To examine how participants felt regarding the experienced environment, the scores for each dimension of the questionnaires were averaged and submitted to a multivariate repeated measures analysis of variance (MANOVA) with factors avatar (2 levels: similar and standard) and session (4 levels) and using the 5 dimensions as measures (realism, possibility, quality, ownership, and self-evaluation).

Finally, Pearson’s *r* was used to test the correlation between critical ratios and the dimension of ownership as measured by the questionnaire.

**Learning of Critical Ratios.** Overall, there was a significant learning effect on the critical ratios,  $F(12, 86) = 6.85, p < .001, \eta^2 = .49$ . This was reflected in the four conditions: slow-similar,  $F(3, 33) = 17.74, p < .001, \eta^2 = .62$ ; fast-similar,  $F(3, 33) = 3.08, p < .05, \eta^2 = .22$ ; slow-standard,  $F(3, 33) = 8.74, p < .001, \eta^2 = .44$ ; and fast-standard,  $F(3, 33) = 3.07, p = .05, \eta^2 = .22$ .

Pairwise comparisons showed significant differences between the first and the last three sessions which did not differ between them (see Figure 4).

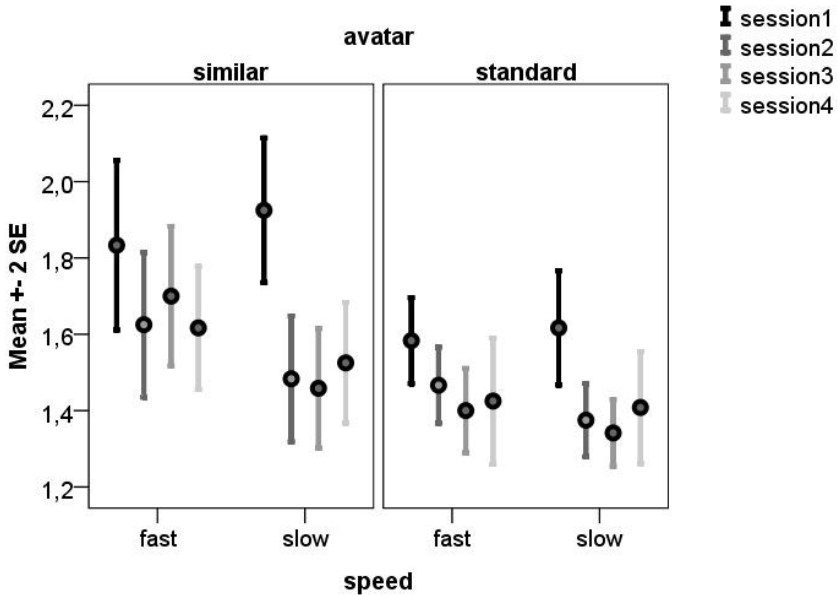
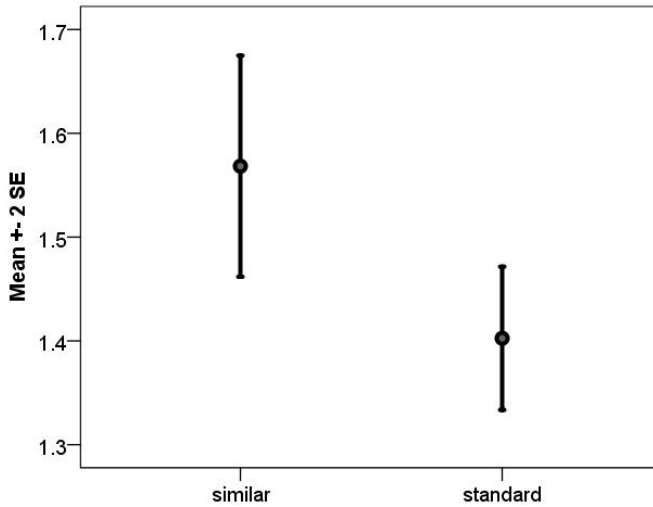


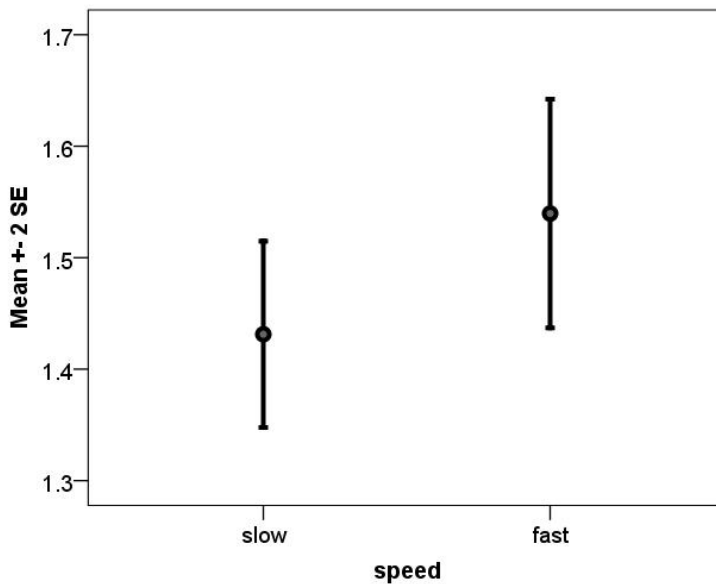
Fig. 4. Average critical ratios for the four conditions over the four sessions

**Critical Ratios.** There was a strong tendency for an effect of critical ratio on avatar,  $F(1, 11) = 4.62, p = .055, \eta^2 = .30$  (Figure 5), which was caused by participants rotating their shoulders at smaller critical ratios when the avatar was standard than when the avatar was similar (standard  $M = 1.40, se = 0.05$ , similar  $M = 1.57, se = 0.07$ ).



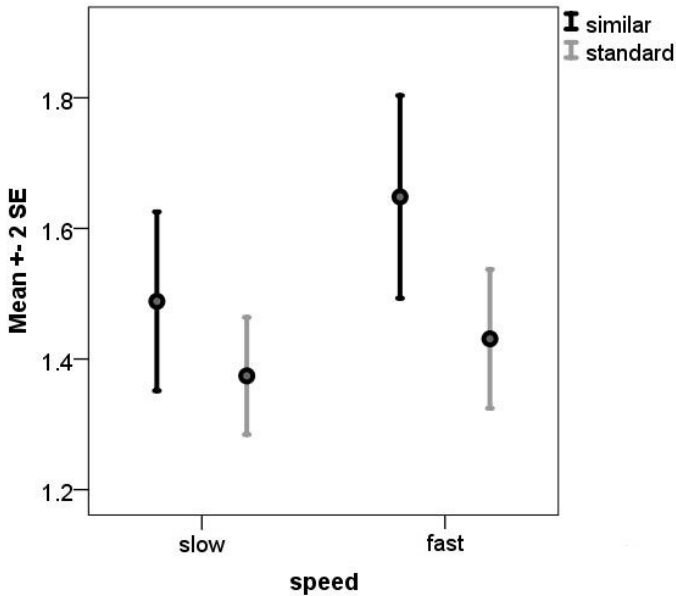
**Fig. 5.** Effect of the avatar on the critical ratios

There was a significant main effect of critical ratio on speed,  $F(1, 11) = 5.13$ ,  $p < .05$ ,  $\eta^2 = .57$  (Figure 6). This was caused by participants rotating their shoulders at smaller critical ratios in the slow condition compared to the fast condition (slow  $M = 1.43$ ,  $se = 0.04$ ; fast  $M = 1.54$ ,  $se = 0.05$ ).



**Fig. 6.** Effect of speed on critical ratios

There was no Speed  $\times$  Avatar interaction,  $F(1, 11) = 2.51$ , however it is noteworthy that the effect of speed was more marked when avatars were similar than when avatars were standard (Figure 7).



**Fig. 7.** Effect of speed condition and avatar on the critical ratios in the four sessions (black bars represent the similar avatar and grey lines represent the standard avatar)

**Duration.** The main effect of duration on speed,  $F(1, 11) = 7.07$ ,  $p < .001$ ,  $\eta^2 = .99$ , was caused by the condition itself (slow  $M = 476.5$ ,  $se = 2.7$ ; fast  $M = 248.3$ ,  $se = 2.2$ ). The main effect of avatar was not statistically significant,  $F(1, 11) = 1.93$ , although participants took slightly shorter in the standard than in the similar condition (standard  $M = 259.3$ ,  $se = 4.06$ , similar  $M = 365.6$ ,  $se = 1.6$ ). There was a significant Speed  $\times$  Avatar interaction,  $F(1, 11) = 19.09$ ,  $p < .001$ ,  $\eta^2 = .63$ . This was because in the slow condition, participants performed slower with similar avatars (slow standard  $M = 468.3$ ,  $se = 5.03$ , slow similar  $M = 484.7$ ,  $se = 2.05$ ), whereas in the fast condition participants performed faster with similar avatars (fast standard  $M = 250.2$ ,  $se = 4.42$ , fast similar  $M = 246.4$ ,  $se = 1.32$ ).

**Collisions.** There were no significant main effects of collisions on speed,  $F(1, 11) = 0.0$ , avatar,  $F(1, 11) = 2.19$ , and no significant interaction effect,  $F(1, 11) = 0.40$ . On average there were 2 collisions on each session across conditions.

**Questionnaire.** Overall, there was no significant main effect of avatar,  $F(2, 7) = 1.83$  or session,  $F(2, 7) = 0.93$ . However, there was a significant Avatar  $\times$  Session interaction,  $F(15, 80) = 1.87$ ,  $p < .05$ ,  $\eta^2 = .24$ . This significant interaction was reflected in



three dimensions: realism,  $F(3, 33) = 4.00, p < .05, \eta^2 = .27$ ; ownership,  $F(3, 33) = 3.93, p < .05, \eta^2 = .26$  and self-evaluation,  $F(3, 33) = 3.17, p < .05, \eta^2 = .22$ . This interaction occurred because feelings of realism, ownership and self-evaluation increased in the similar condition and decreased in the standard condition (Figure 8).

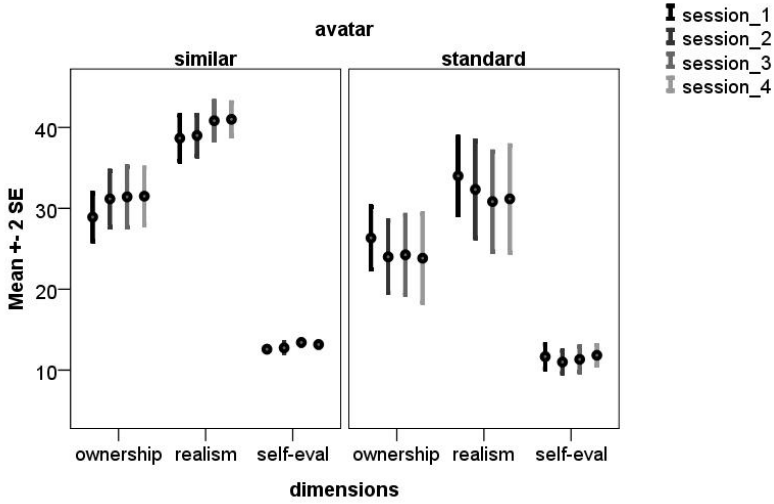


Fig. 8. Questionnaires results for the three dimensions over sessions

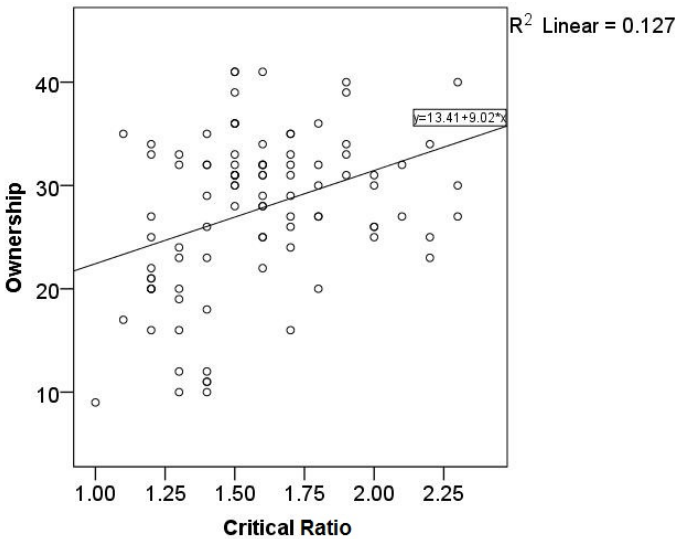


Fig. 9. Scatter plot and Pearson's r

**Ownership and Critical Ratios.** Overall, there was a small, positive correlation between feelings of ownership and critical ratios,  $r = 0.36$ ,  $n = 96$ ,  $p < 0.05$  indicating that increases in one variable were accompanied by increases in the other variable (Figure 9).

## 4 Discussion

The main objective of this research was to study whether the dynamics and the morphology of an avatar would reflect on the feelings of telepresence and ownership of the participant. Also, to know whether affordances were used in a virtual environment the same way they are in a real environment. In order to perform the experiment, a prototype named ATTAVE was developed. ATTAVE possess a virtual environment where the participant's avatar performs the experimental task. The user controlled the avatar through a natural motion capture carried out by a Kinect NUI. There were two avatar conditions to test: one morphologically proportional to the user and that replicated his/her movements; and another with an avatar that was identical for every participant, and exhibited limited mobility, as it could only rotate upon himself and step sideways.

The results show that participants adjusted to the virtual environment after taking their first session. This learning effect is only significant in the first session, and the other three sessions were similar to each other in a same condition. The learning effect that occurred in the first session indicates that participants learned to use appropriate information, provided by the virtual environment, to solve the problem of passing through apertures. It is usual in Human-Machine Interaction (HMI) that users take time to adapt to the artifact. For instance, Peters et al. [28] showed that six trials on a teleoperated robotic arm are necessary to have an accurate representation of the task. On the contrary to a simple tool held in hand, the remote or virtual interaction is not straightforward and involves a specific design of the artifact to be ergonomically adapted to the human user [29]. After learning, the critical ratios obtained in the virtual environment (1.4 and 1.57) were very similar to those in the real environment (1.4 reported by Warren and Whang [18]), which demonstrates that people perceive similar body-scaled affordances in the virtual environment as in real environments.

The effect of the different avatars was significant. Participants with the standard avatar had a smaller critical ratio than the participants with the similar avatar (1.4 vs. 1.57). It means that participants who controlled a similar avatar tended to keep a greater safety margin between the body and wall, compared with a standard avatar. This result can suggest a higher feeling of agency for the similar avatar by the individual, who may care about his/her alter-ego and therefore make sure it does not collide with the wall. In contrast, participant seemed less concerned regarding the standard avatar. This interpretation is supported by the analysis of the correlation between the ownership questionnaire and the critical ratios. There was a positive correlation between these two variables, albeit small, which suggests that the rise in the critical ratios is related with an increase in the feelings of ownership.

The higher critical ratios in the similar condition compared to the standard condition could also be explained by distraction because the similar condition involved seeing more of their own movements in the avatar's movements. The amount movements may have led the participants to attend to something other than the shoulders and the apertures. The movements of the standard avatar were much more restricted, which decreased the amount of cues available. In the standard condition, the participant's body controlled the avatar as if it were a simple joystick. Consequently, this condition can be interpreted in terms of a classical remote control of a teleoperated machine. In a study carried out by Moore et al. [30], the user controlled a robot and supervises the environment by means of a camera on top of the robot. The operator's task was to judge whether or not the robot could pass through apertures of various sizes. The results indicate that the participants judged the robot could pass even when it could not. Authors argued that the results might be influenced by structural and morphological aspects. The same effect could explain the results of the experiment presented here. The fact that the individual underestimates the aperture width could signify that he/she considers the standard avatar as a machine-like character and, consequently, he/she can hardly be engaged in an ownership process with it.

In addition, one aspect that can increase the immersiveness in a virtual environment is the type of interaction between human and machine. If the interaction is done through a classic controller such as a mouse or joystick, the users need to learn the mapping between their own movement and its consequence in the virtual world [31, 32]. On the contrary, the mapping is facilitated if the interaction is done through a full motion capture. Research has shown that natural user interfaces, in which users can recognize their own movements in the virtual environment, are more immersive [33, 34]. It could explain why the similar avatar condition seems to bring a higher feeling of ownership compared with the standard condition.

Analyzing the effect of speed on the critical ratios, it is observable that in both avatar conditions the critical ratio is larger when the speed is higher. This happens because at higher speeds people leave larger safety margins. In the real world, when an individual is confronted with an aperture, he/she will reduce his/her speed in order to fit through without colliding. In this project, the only way to decrease the speed was by rotating the shoulders, which resulted in a higher critical ratio. The relation between speed and accuracy is well-known in the area of motor control and was described for the first time in Fitts' law [35]. The fact that the participants of this experiment reproduced this natural motor control adaptation suggests their immersion into the virtual environment (ATTAVE).

Finally, the questionnaire showed that the similar avatar elicited an increasing feeling of ownership and realism over the four sessions. In contrast, the standard avatar caused a decreasing feeling of ownership and realism from session to session. This result means that an agency process in which the participants consider the similar avatar as a natural extension of themselves seems to happen and to increase through the interaction with the avatar. In contrast, with the standard avatar, participants may have become bored due to the avatar's movement not being as diverse as their own. This lack of biological motion could lead the participants to act as if they were controlling a machine instead of a virtual representation of themselves. Overall,

considering the questionnaire results and the affordances described through the critical ratios, it seems that an avatar with natural movements and tailored the morphology of the user can significantly enhance the feelings of body ownership.

## References

1. Minsky, M.: Telepresence. *Omni* 2, 45–51 (1980)
2. Slater, M., Usoh, M., Steed, A.: Depth of presence in virtual environments. *Presence* 3, 130–144 (1994)
3. Rybarczyk, Y., Mestre, D.: Effect of visuo-manual configuration on a telerobot integration into the body schema. *Le Travail Humain* 76, 181–204 (2013)
4. Rybarczyk, Y., Hoppenot, P., Colle, E., Mestre, D.: Sensori-motor appropriation of an artefact: A neuroscientific approach. In: Inaki, M. (ed.) *Human Machine Interaction - Getting Closer*, pp. 187–212. InTech, Rijeka (2012)
5. Sumioka, H., Nishio, S., Ishiguro, H.: Teleoperated android for mediated communication: Body ownership, personality distortion, and minimal human design. *Social Robotic Telepresence* 32 (2012)
6. Botvinick, M., Cohen, J.: Rubber hands ‘feel’ touch that eyes see. *Nature* 391, 756–756 (1998)
7. Yuan, Y., Steed, A.: Is the rubber hand illusion induced by immersive virtual reality? In: *Virtual Reality Conference*, pp. 95–102. IEEE Press (2010)
8. Tsakiris, M., Prabhu, G., Haggard, P.: Having a body versus moving your body: How agency structures body-ownership. *Consciousness and Cognition* 15, 423–432 (2006)
9. Ehrsson, H.H., Spence, C., Passingham, R.E.: That’s my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science* 305, 875–877 (2004)
10. Petkova, V.I., Ehrsson, H.H.: If I were you: Perceptual illusion of body swapping. *PLoS One* 3, e3832 (2008)
11. Maravita, A., Iriki, A.: Tools for the body (schema). *Trends in Cognitive Sciences* 8, 79–86 (2004)
12. Iriki, A., Tanaka, M., Iwamura, Y.: Coding of modified body schema during tool use by macaque postcentral neurons. *Neuroreport* 7, 2325–2330 (1996)
13. Maselli, A., Slater, M.: The building blocks of the full body ownership illusion. *Frontiers in Human Neuroscience* 7, 83 (2013)
14. Armel, K.C., Ramachandran, V.S.: Projecting sensations to external objects: Evidence from skin conductance response. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270, 1499–1506 (2003)
15. Slater, M., Perez-Marcos, D., Ehrsson, H.H., Sanchez-Vives, M.V.: Inducing illusory ownership of a virtual body. *Frontiers in Neuroscience* 3, 214 (2009)
16. González-Franco, M., Peck, T.C., Slater, M.: Virtual Embodiment Elicits a Mu Rhythm ERD When the Virtual Hand is Threatened. In: 8th International Brain Research Organisation, Congress of Neuroscience (2011)
17. Gibson, J.J.: *The ecological approach to visual perception*. Houghton Mifflin, Boston (1979)
18. Warren, W.H., Whang, S.: Visual guidance of walking through apertures: Body-scaled information for affordances. *Journal of Experimental Psychology: Human Perception and Performance* 13, 371–383 (1987)

19. De Oliveira, R.F., Damisch, L., Hossner, E.J., Oudejans, R.D., Raab, M., Volz, K.G., Williams, A.M.: The bidirectional links between decision making, perception, and action. *Progress in Brain Research* 174, 85–93 (2009)
20. Mark, L.S.: Eyeheight-scaled information about affordances: A study of sitting and stair climbing. *Journal of Experimental Psychology: Human Perception and Performance* 13, 361–370 (1987)
21. Esteves, P.T., De Oliveira, R.F., Araújo, D.: Posture-related affordances guide attack in basketball. *Psychology of Sport and Exercise* 12, 639–644 (2011)
22. Warren, W.H.: Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance* 10, 683–703 (1984)
23. Meadows, M.S.: *I, avatar*. New Riders Press, Berkley (2008)
24. Castronova, E.: *Theory of the Avatar*. Cesifo (2003)
25. Tsakiris, M., Haggard, P.: The rubber hand illusion revisited: Visuotactile integration and self-attribution. *Journal of Experimental Psychology: Human Perception and Performance* 31, 80–91 (2005)
26. Kalckert, A., Ehrsson, H.H.: Moving a rubber hand that feels like your own: A dissociation of ownership and agency. *Frontiers in Human Neuroscience* 6, 40 (2012)
27. Witmer, B.J., Jerome, C.J., Singer, M.J.: The factor structure of the Presence questionnaire. *Presence* 14, 298–312 (2005)
28. Peters, R.A., Campbell, C.L., Bluethmann, W.J., Huber, E.: Robonaut task learning through teleoperation. In: *IEEE International Conference on Robotics and Automation*, pp. 2806–2811. IEEE Press (2003)
29. Rybarczyk, Y., Mestre, D.: Effect of temporal organization of the visuo-locomotor coupling on the predictive steering. *Frontiers in Psychology* 3, 239 (2012)
30. Moore, K.S., Gomer, J.A., Pagano, C.C., Moore, D.D.: Perception of robot passability with direct line of sight and teleoperation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 51, 557–570 (2009)
31. De Oliveira, R.F., Wann, J.P.: Driving skills of young adults with developmental coordination disorder: Regulating speed and coping with distraction. *Research in Developmental Disabilities* 32, 1301–1308 (2011)
32. Wise, S.P., Murray, E.A.: Arbitrary associations between antecedents and actions. *Trends in Neurosciences* 23, 271–276 (2000)
33. Bruder, G., Steinicke, F., Hinrichs, K.H.: Arch-explore: A natural user interface for immersive architectural walkthroughs. In: *IEEE Symposium on 3D User Interfaces*, pp. 75–82. IEEE Press (2009)
34. Francese, R., Passero, I., Tortora, G.: Wiimote and Kinect: Gestural user interfaces add a natural third dimension to HCI. In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 116–123. ACM Press (2012)
35. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 381–391 (1954)

# Reactive Statistical Mapping: Towards the Sketching of Performative Control with Data

Nicolas d’Alessandro<sup>1</sup>, Joëlle Tilmanne<sup>1</sup>, Maria Astrinaki<sup>1</sup>,  
Thomas Hueber<sup>2</sup>, Rasmus Dall<sup>3</sup>, Thierry Ravet<sup>1</sup>, Alexis Moinet<sup>1</sup>,  
Huseyin Cakmak<sup>1</sup>, Onur Babacan<sup>1</sup>, Adela Barbulescu<sup>2</sup>, Valentin Parfait<sup>1</sup>,  
Victor Huguenin<sup>1</sup>, Emine Sümeyye Kalaycı<sup>1</sup>, and Qiong Hu<sup>3</sup>

<sup>1</sup> Numediart Institute for New Media Art Technology, University of Mons, Belgium  
{nicolas.dalessandro,joelle.tilmanne,maria.astrinaki,thierry.ravet,  
alexis.moinet,huseyin.cakmak,onur.babacan}@umons.ac.be

<sup>2</sup> GIPSA-lab, UMR 5216/CNRS/INP/UJF/Stendhal University, Grenoble, France  
{thomas.hueber, adela.barbulescu}@gipsa-lab.grenoble-inp.fr

<sup>3</sup> Centre for Speech Technology Research, University of Edinburgh, Scotland, UK  
r.dall@sms.ed.ac.uk, qiong.hu@ed.ac.uk  
<http://www.numediart.org/hmapper>

**Abstract.** This paper presents the results of our participation to the ninth eNTERFACE workshop on multimodal user interfaces. Our target for this workshop was to bring some technologies currently used in speech recognition and synthesis to a new level, i.e. being the core of a new HMM-based mapping system. The idea of statistical mapping has been investigated, more precisely how to use Gaussian Mixture Models and Hidden Markov Models for realtime and reactive generation of new trajectories from inputted labels and for realtime regression in a continuous-to-continuous use case. As a result, we have developed several proofs of concept, including an incremental speech synthesiser, a software for exploring stylistic spaces for gait and facial motion in realtime, a reactive audiovisual laughter and a prototype demonstrating the realtime reconstruction of lower body gait motion strictly from upper body motion, with conservation of the stylistic properties. This project has been the opportunity to formalise HMM-based mapping, integrate various of these innovations into the MAGE library and explore the development of a realtime gesture recognition tool.

**Keywords:** Statistical Modelling, Hidden Markov Models, Motion Capture, Speech, Singing, Laughter, Realtime Systems, Mapping.

## 1 Introduction

The simulation of human communication modalities, such as speech, facial expression or body motion, has always been led by the challenge of making the virtual character look “more realistic”. Behind this idea of realness, there are inherent properties that listeners or viewers have learnt to expect and track with great accuracy. Empirical studies, such as the Mori’s vision for robotics [1], tend

to demonstrate that this quest for human likelihood is highly non-linear, encountering a well-known phenomenon called the *uncanny valley*, i.e. an unexpected shift from empathy to revulsion when the response of the virtual character has “*approached, but failed to attain, a lifelike appearance*” [2].

## 1.1 A Content-Oriented Approach of Expressivity

For the last decades, research fields like sound synthesis, computer graphics or computer animation have faced this issue of the uncanny valley in their own ways. However common trends can be highlighted. The primary goal in producing artificial human modalities has always been to “preserve the message”, i.e. what is heard or seen is at least understood correctly. In speech synthesis, this is called *intelligibility* [3]. We can transpose this property to images or motion, as it refers to the readability of what is generated: a smile, a step, a laughter sound, etc. Later the target has evolved to “make it look more natural”. This trend of *naturalness* is what has brought nearly everybody in these fields to use recordings of actual human performances, thus moving beyond explicit rule-based modelling. We can retrieve this concept in the development of non-uniform unit selection in speech synthesis [4], giga-sampling in music production [5], high-resolution face scanning [6] or full-body motion capture [7] in animation.

Nowadays the target is to bring *expressivity* and *liveliness* to these virtual characters. This idea goes further than naturalness as it expects the virtual character to display a wide range of convincing emotions, either automatically or by means of some authoring. For at least a decade, the typical approach to this problem has been to extend the amount and variability of the recorded data, hoping to gather enough utterances of the expected emotions so as to encompass what we perceive as human. If this idea is meaningful, one might conjecture that the way of representing and using this massive amount of data is not very lively nor flexible. For instance, non-uniform unit selection in speech synthesis concatenates very long utterances from the original recordings [4] and static or dynamic postures in animation are often blended from original sequences without a deep understanding of the production mechanisms.

## 1.2 Performative Control and Machine Learning

Although the use of a large amount of data has clearly improved some aspects of virtual human-like character animation<sup>1</sup>, we could argue from the current results that this approach on its own has not been able to fully climb the uncanny valley. Indeed speech synthesis and face/body animation resulting from monolithic transformations of original recordings keep having something inappropriate and

---

<sup>1</sup> Our approach is really transversal to all the virtual character’s modalities: voice, face and body. Therefore we tend to use the terms “rendering” and “animation” for both sounds and visuals. This distinction is rarely done for sound, as “synthesis” is often used for the whole process. Though a similar nuance can be found in speech, when respectively referring to segmental and suprasegmental qualities [3].

confusing [8]. In this research, we think that user interaction has a great role to play in how a given virtual human-like character will be considered more expressive and lively. Indeed we agree that expressivity is a matter of displaying a great variability in the rendered output, but we think that these variations need to be contextually relevant. This “context” is a sum of elements surrounding the virtual character at a given time and user interaction has a big impact on how the rendering system should behave. Our way of reading the Mori’s law is that a significant part of the perceived humanness sits in very subtle details. In this work, we will focus on how the animation trajectories can reactively adapt to the user interaction on the very short term, i.e. within milliseconds. We think that there is a lack in the literature about how these trajectories should react according to very short-term gestures, as most of the research is focusing on a larger time window and the overall discussion of dialog systems.

Our approach towards bringing together large multimodal datasets and short-term user interaction is to use machine learning. There is already a very large body of literature on applying machine learning techniques in the fields of gesture or voice recognition [9], gesture or voice synthesis [10], gesture or voice conversion [11], and what is called *implicit mapping*, i.e. the use of statistical layers as a way of connecting inputs to outputs in interactive systems [12]. Multimodal animation of virtual characters brings two main constraints to take into account when it comes to statistical modelling. On the one hand, dimensionality of the data is very big (see Section 3 for details). On the other hand, we aim at creating animation *trajectories*, which means that the temporal quality of the generated outputs is crucial. Besides these specific animation-related constraints, we want to enable short-term interaction with the virtual character. Therefore the way of handling statistics in our system requires to be fully *realtime*.

### 1.3 A New Framework for GMM/HMM-Based Mapping

In this project, we have decided to investigate the use of Gaussian Mixture Modelling (GMM) and Hidden Markov Modelling (HMM) as statistical tools to abstract big collections of multimodal data, use such knowledge to animate virtual characters in realtime and enable various kinds of user interactions to happen. GMM/HMM offers great ability to cover complex feature spaces and HMM is designed for taking care of the temporal structure of trajectories to be rendered. Moreover the development of GMM/HMM-based machine learning techniques has been greatly boosted by recent advances in speech technology research. Particularly the idea of HMM-based speech synthesis has emerged in the last decade [10]. A synthesis system called HTS has become a reference in this field [13]. There are many pieces of innovative research surrounding GMM/HMM-based generation. Three of them have particularly encouraged our team to envision a new GMM/HMM-based framework for virtual character animation:

- the adaptation of the speech algorithms to motion capture data [14,15];
- the modification of the core generation algorithms to be fully realtime [16];
- the integration of mapping functions inside the HMM framework [17,18].



As a result, we have decided to create a new framework that accepts and decodes user interaction gestures in realtime, finds the appropriate statistical context to apply mapping strategies and is able to generate new trajectories to partially or totally animate a virtual character. In this work, this prototype framework has been tested for a great amount of use cases, encompassing many modalities: speech, singing, laughter, face and body motion.

## 1.4 Outline of the Paper

In this paper, we will first present a more detailed background theory on the statistical models that are used in our system (cf. Section 2). In Section 3 we describe the particularities of the datasets that we have been using: speech, laughter, singing, facial and gait motion. Section 4 explains the various use cases in which we enabled the realtime trajectory generation. Section 5 focuses on use cases where mapping is the fundamental feature. Further details on the architecture are given in Section 6. Finally we conclude in Section 7.

## 2 Statistical Feature Mapping: Theoretical Aspects of GMM-Based and HMM-Based Mapping Techniques

The research described in this paper relies on a very specific use of machine learning techniques based on Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). The aim of this Section is to describe a set of background theories that are necessary to understand the following Sections. In the next parts, we first recall the theoretical context of *statistical feature mapping*. Then we give an overview on how to turn GMM and HMM into mapping layers. We also give more details on how to generate the HMM-based trajectories in realtime, as it is required for building an interactive system.

### 2.1 Statistical Feature Mapping

The problem of *feature mapping* refers to the prediction of a vector of target variables  $\hat{\mathbf{y}}$  (also called target *features*), from the observation of an unseen vector of input variables  $\mathbf{x}$ . This problem can be divided into three categories:

- *regression*, also referred to as *continuous mapping*, i.e. the case where both  $\mathbf{y}$  and  $\mathbf{x}$  will comprise continuous variables;
- *classification*, i.e. the case where  $\mathbf{y}$  will represent a discrete set of class labels while  $\mathbf{x}$  will comprise continuous variables;
- *generation*, also referred to as *synthesis*, i.e. the case where  $\mathbf{x}$  will represent class labels and  $\mathbf{y}$  the continuous variables we want to estimate.

The feature mapping problem can be viewed from a probabilistic viewpoint. It consists in finding the vector of target variables  $\hat{\mathbf{y}}$  which maximises the conditional probability  $p(\mathbf{y}|\mathbf{x})$  of  $\mathbf{y}$  given  $\mathbf{x}$ , such as:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} \{ p(\mathbf{y}|\mathbf{x}) \} \quad (1)$$

In *supervised* machine learning, this conditional probability is estimated during the training phase from a dataset  $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$  comprising  $N$  observations of  $\mathbf{x}$  together with the corresponding observations of the values of  $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]$ . Two types of approaches can be envisioned to estimate this conditional probability: the discriminative or the generative approach.

In a *discriminative* approach, this conditional probability distribution (also referred to as the posterior probability distribution) is estimated directly from the training data. In other words, such an approach provides a model only for the target variables conditional on the observed variables and does not provide a complete probabilistic model of all the variables (observed and hidden ones). Examples of discriminative models include Linear Discriminant Analysis (LDA), Conditional Random Fields (CRF), Artificial Neural Networks (ANN), etc.

In a *generative* approach, the conditional probability is not estimated directly, but derived from the joint probability distribution  $p(\mathbf{x}, \mathbf{y})$  using Bayes’ theorem:

$$p(\mathbf{x}, \mathbf{y}) = \mathbf{p}(\mathbf{y}|\mathbf{x})\mathbf{p}(\mathbf{x}) = \mathbf{p}(\mathbf{x}|\mathbf{y})\mathbf{p}(\mathbf{y}) \quad (2)$$

where  $p(\mathbf{x}|\mathbf{y})$  is the likelihood function,  $p(\mathbf{y})$  the prior probability – i.e. the probability of  $\mathbf{y}$  *before* seeing  $\mathbf{x}$  – and  $p(\mathbf{x})$  is usually viewed as a normalisation constant. Note that the joint probability  $p(\mathbf{x}, \mathbf{y})$  can be either modelled explicitly or implicitly via the separate estimation of the likelihood function and the prior probability. Examples of generative models includes GMMs and HMMs.

**Using Generative Models.** Generative and discriminative approaches have their own advantages and drawbacks. An extensive discussion about which approaches would be more suitable for addressing a specific application is far beyond the scope of this paper. However we focused on generative models for two main reasons.

The first advantage of the generative approach is that the inclusion of *prior knowledge* arises naturally. This may be extremely convenient to address problems, considered as ill-posed, for which there is no clear one-to-one mapping between input and target feature spaces. Furthermore, as prior probabilities do not depend on the input observation  $\mathbf{x}$ , they may be estimated on a much larger dataset than the training set. Prior knowledges can be used to constrain the mapping process to generate acceptable outputs (as in speech recognition, where a language model gives the probability of having a specific sequence of words in a specific language, independently from the observed acoustic signal).

The second advantage is its *flexibility*. Estimating the joint probability distribution over input and target variables  $p(\mathbf{x}, \mathbf{y})$  allows to address several mapping problems with the same model. As an example, let us consider a mapping problem between a continuous feature space  $\mathbf{x}$  and a discrete space of  $k$  classes  $C_k$ . The same approach which aims at estimating  $p(\mathbf{x}, \mathbf{C}_k)$  could be used to address both the related classification problem by deriving the conditional probability  $p(C_k|\mathbf{x})$ , and the symmetrical problem of trajectory generation, by sampling the conditional probability  $p(\mathbf{x}|\mathbf{C}_k)$  – i.e. the likelihood function.

However we have to keep in mind that generative approaches are known to require a lot of training data. In fact, the dimensionality of both input  $\mathbf{x}$  and target observations  $\mathbf{y}$  may be high, and consequently a large training set is needed in order to be able to determine  $p(\mathbf{x}, \mathbf{y})$  accurately.

In this project, we focused on two generative approaches, which are GMMs and HMMs. The theoretical aspects of these two techniques are mainly presented in the context of *regression*, plus a specific focus on *realtime generation* and the MAGE framework. Indeed more general aspects of GMM/HMM-based classification [9] and generation [10] have already been extensively described.

From now on, sequences of input and target feature vectors are noted respectively  $\mathbf{x}$  and  $\mathbf{y}$ , and are defined as:  $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_t, \dots, \mathbf{x}_T]$  and  $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_t, \dots, \mathbf{y}_T]$ , where  $\mathbf{x}_t$  are  $\mathbf{y}_t$  are respectively  $D_x$  and  $D_y$  dimensional vectors of input and target features observed at time  $t$  ( $T$  is the sequence length).

## 2.2 GMM-Based Mapping

During the training phase, the joint probability density function (pdf) of input and target features is modelled by a GMM such as:

$$p(\mathbf{z}|\lambda) = p(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^M \alpha_m N(\mathbf{z}, \mu_m, \Sigma_m) \quad (3)$$

with

$$\mathbf{z} = [\mathbf{x}, \mathbf{y}], \quad \mu_m = \begin{bmatrix} \mu_m^{\mathbf{x}} \\ \mu_m^{\mathbf{y}} \end{bmatrix}, \quad \Sigma_m = \begin{bmatrix} \Sigma_m^{\mathbf{xx}} & \Sigma_m^{\mathbf{xy}} \\ \Sigma_m^{\mathbf{yx}} & \Sigma_m^{\mathbf{yy}} \end{bmatrix} \quad (4)$$

where  $\lambda$  is the parameter set of the model,  $N(\cdot, \mu, \Sigma)$  is a normal distribution with mean  $\mu$  and covariance matrix  $\Sigma$ ,  $M$  is the number of mixture components, and  $\alpha_m$  is the weight associated with the  $m^{\text{th}}$  mixture component (prior probabilities). Given a training dataset of input and target feature vectors, the parameters of a GMM (weights, mean vectors and covariance matrices for each component) are usually estimated using the expectation-maximisation (EM) algorithm. The initial clustering of the training set can usually be obtained using the k-means algorithm.

In the mapping phase, a conditional pdf  $p(\mathbf{y}_t|\mathbf{x}_t, \lambda)$  is derived, for each frame  $t$ , from the joint pdf  $p(\mathbf{x}_t, \mathbf{y}_t)$  estimated during training, such as described in Eq. 5 to 9. The mathematical basis of this derivation can be found in [19].

$$p(\mathbf{y}_t|\mathbf{x}_t, \lambda^{[\mathbf{xy}]}) = \sum_{m=1}^M p(c_m|\mathbf{x}_t) p(\mathbf{y}_t|\mathbf{x}_t, m, \lambda^{[\mathbf{xy}]}) \quad (5)$$

where  $\lambda$  is the model parameter set. The posterior probability  $P(c_m|\mathbf{x}_t)$  of the class  $c_m$  given the input vector  $\mathbf{x}_t$ , and the mean and covariance of the class-dependent conditional probability  $P(\mathbf{y}_t|\mathbf{x}_t, m, \lambda^{[\mathbf{xy}]})$  are defined as:

$$p(c_m|\mathbf{x}_t) = \frac{\alpha_m N(\mathbf{x}_t, \mu_m^{\mathbf{x}}, \Sigma_m^{\mathbf{x}\mathbf{x}})}{\sum_{p=1}^M \alpha_p N(\mathbf{x}, \mu_p^{\mathbf{x}}, \Sigma_p^{\mathbf{x}\mathbf{x}})} \quad (6)$$

and

$$p(\mathbf{y}_t|\mathbf{x}_t, m, \lambda^{[\mathbf{x}\mathbf{y}]}) = N(\mathbf{y}_t, E_{(m,t)}, D_{(m,t)}) \quad (7)$$

with

$$E_{(m,t)} = \mu_m^{\mathbf{y}} + \Sigma_m^{\mathbf{y}\mathbf{x}} \Sigma_m^{\mathbf{x}\mathbf{x}}^{-1} (\mathbf{x}_t - \mu_m^{\mathbf{x}}) \quad (8)$$

$$D_{(m,t)} = \Sigma_m^{\mathbf{y}\mathbf{y}} - \Sigma_m^{\mathbf{y}\mathbf{x}} \Sigma_m^{\mathbf{x}\mathbf{x}}^{-1} \Sigma_m^{\mathbf{x}\mathbf{y}} \quad (9)$$

Two approaches can then be envisioned to address a regression problem with a GMM. In the first one, referred here to as the *MMSE-GMR* for “Gaussian Mixture Regression based on the Minimum Mean Square Error Criterion”, the target feature vector  $\hat{\mathbf{y}}_t$  estimated from the given source vector  $\mathbf{x}_t$  observed at time  $t$ , is defined as  $\hat{\mathbf{y}}_t = E[\mathbf{y}_t|\mathbf{x}_t]$  (where  $E$  means expectation) and can be calculated by solving Eq. 10. In particular, this approach has been described in [20] and [21] in the context of statistical voice conversion.

$$\hat{\mathbf{y}}_t = \sum_{m=1}^M p(c_m|\mathbf{x}_t) E_{m,t}^{\mathbf{y}} \quad (10)$$

The second approach, proposed by Toda et al. [22], is here referred to as *MLE-GMR*, for “Gaussian Mixture Regression based on Maximum Likelihood Criterion”. The target feature vector  $\hat{\mathbf{y}}_t$  is defined as the one which maximises the likelihood function such as:

$$\hat{\mathbf{y}}_t = \arg \max_{\mathbf{y}_t} \{ p(\mathbf{y}_t|\mathbf{x}_t, \lambda^{[\mathbf{x}\mathbf{y}]}) \} \quad (11)$$

and can be estimated by solving Eq. 12:

$$\hat{\mathbf{y}}_t = \left[ \sum_{m=1}^M p(c_m|\mathbf{x}_t) D_{(m,t)}^{\mathbf{y}} \right]^{-1} \left[ \sum_{m=1}^M p(c_m|\mathbf{x}_t) D_{(m,t)}^{\mathbf{y}} E_{(m,t)}^{\mathbf{y}} \right] \quad (12)$$

**Trajectory GMM.** The MLE-GMR approach is commonly combined with a constraint on the smoothness of the predicted trajectories. The GMM is then referred to as a *trajectory GMM*. In that case, each target feature vector  $\mathbf{y}_t$  of the training set is augmented by its  $N$ -order derivatives such as  $\mathbf{Y}_t = [\mathbf{y}_t \Delta \mathbf{y}_t]$  (the method is here presented with  $N = 1$ ). The joint probability distribution  $p(\mathbf{Y}, \mathbf{x})$  is modelled similarly to the MLE-GMR approach. However the mapping process is done differently. The sequence of target feature vectors is not

estimated on a frame-by-frame basis as in previous approaches, but rather “all-at-once”. The estimated sequence of target feature vectors is defined as the one that maximises the likelihood of the model, with respect to the continuity of its first  $N$  derivatives:

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}} \{ p(\mathbf{Y}|\mathbf{x}, \lambda^{[\mathbf{x}^{\mathbf{Y}}]}) \} \quad (13)$$

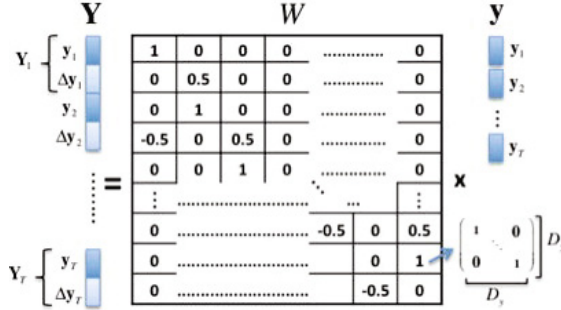
which can be estimated by solving the closed-form equation in Eq. 14:

$$\hat{\mathbf{y}} = (W^T D_{\hat{m}}^{-1} W)^{-1} W^T D_{\hat{m}}^{-1} E_{\hat{m}} \quad (14)$$

where  $W$  is a  $[2D_x T \times D_y T]$  matrix representing the relationship between static and dynamic feature vectors (Fig. 1) and  $\hat{m} = [\hat{m}_1, \dots, \hat{m}_t, \dots, \hat{m}_T]$  is the sub-optimum sequence of mixture components defined as:

$$\hat{m} = \arg \max_m \{ p(c_m | \mathbf{x}, \lambda) \} \quad (15)$$

and determined using the Viterbi algorithm (in our experiment, and similarly to what was reported in [22], similar results were obtained using a forward-backward approach which takes into account in a probabilistic manner the contributions of all mixture components).



**Fig. 1.**  $W$  is a  $[2D_x T \times D_y T]$  matrix representing the relationship between static and dynamic feature vectors. It is used in the computation of output trajectories of Eq. 14.

## 2.3 Realtime Trajectory Generation

If the “all-at-once” approach suggested by Eq. 14 can guarantee smooth trajectories, it prevents these trajectories to be generated in realtime, and therefore the system to be interactive. In Section 1, we have claimed that interactivity was a required property for our system, in order to produce expressive virtual characters. For the last few years, Astrinaki et al. have worked to find a new trade-off between the smoothness and the system-wise reactivity of generated

trajectories [16,17]. The idea of finding the Maximum Likelihood (MLE) over the whole sequence as required to fill the whole  $W$  matrix in Eq. 14 has been replaced by finding a sub-optimal version of the ML on a sliding window. This new algorithm, called *Short-Term Maximum Likelihood Parameter Generation (ST-MLPG)*, enables the trajectory generation process to start when the system receives the first class label and not at the end of a full sequence of class labels.

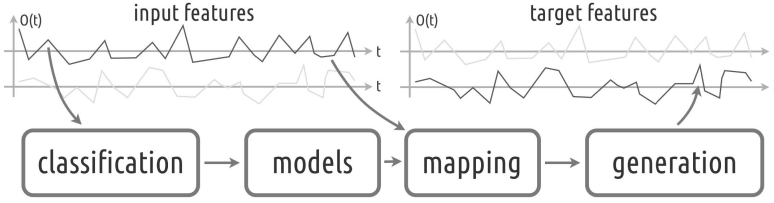
The ST-MLPG is a key feature of the open-source library called MAGE [23]. The MAGE software actually enables to create interactive systems that generate smooth trajectories in realtime, which is the main reason why this software is the trajectory generation engine in this project. However the MAGE library was very tied to the HMM-based approach – explained in the next part – and the HTS architecture when we started the project, which explains why we have significantly modified its architecture along the way (see 6 for details).

## 2.4 HMM-Based Mapping

Hidden Markov Modelling has been used for a long time in *temporal pattern recognition*, as for instance in automatic speech recognition (ASR), handwriting, or gesture recognition. More recently HMM has also been successfully used for parameter generation, such as in HMM-based speech synthesis [10]. HMM can be considered as a generalisation of GMM where the latent variables – i.e. the hidden states – are derived from a Markov process rather than being independent from each other. Latent variables are controlling the mixture component to be selected for each observation. As HMM aims at representing explicitly the temporal evolution of features, it is adapted to model data that can be clustered not only by its distribution but also by its temporal evolution.

Due to this temporal structure, achieving a regression task (or mapping) with a HMM requires a more complex framework. The algorithm proposed in [18] combines a HMM-based classification and a HMM-based parameter generation, from the same trained models. The mapping operation can therefore happen within the HMM currently in use and “passed” from the classification to the generation operations. This process is illustrated in Fig. 2. We traditionally assume that sequences of feature vectors which constitutes the training set are temporally segmented and labeled. This segmentation can be obtained by annotating the data, either manually, or by using an initial set of already-trained HMMs and a forced-alignment procedure.

In the training phase, sequences of input and target feature vectors (completed with their first  $N$  derivatives) are modelled jointly, for each class, by a “full-covariance” HMM, i.e. an HMM for which the emission probability distribution is modelled, for each state  $q$ , by a normal distribution with a full-covariance matrix, as defined by Eq. 3 (with  $M = 1$ ). After initialisation, models are typically trained using the following standard procedure: models are first trained separately, using the standard Baum-Welch re-estimation algorithm and then processed simultaneously, using an embedded training strategy. Since input/target features are often sensitive to context effects (for instance, co-articulation and anticipation in speech), context-dependency is often introduced in the modelling.



**Fig. 2.** Summary of the framework required to achieve a HMM-based mapping: HMM-based classification of input features, query of models based on obtained class labels, mapping routine from models and input features and generation of target features

Context-dependent models are created by adding information about left and right contexts to the initial models. A tree-based state-tying technique is then eventually used to tackle the problem of data sparsity (context-dependent models having only a few occurrences in the training dataset).

In the mapping phase, the sequence of target feature vectors  $\hat{\mathbf{y}}$  is estimated from the sequence of input feature vectors  $\mathbf{x}$  such as:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} \{ p(\mathbf{y}|\mathbf{x}) \} \quad (16)$$

with

$$p(\mathbf{y}|\mathbf{x}, \lambda) = p(\mathbf{Y}|\lambda, q, \mathbf{x}) \cdot p(\lambda, q|\mathbf{x}) \quad (17)$$

with  $Y = W\mathbf{y}$  (see Fig. 1),  $\lambda$  the parameters set of the HMM and  $q$  the HMM state sequence. Eq. 17 is the product of two conditional probability terms which can be maximised separately:

1.  $p(\lambda, q|\mathbf{x})$  which is related to the *classification stage* which aims at estimating the most likely HMM (or sequence of concatenated HMMs)  $\hat{\lambda}$  with the corresponding sequence of states  $\hat{q}$  such as:

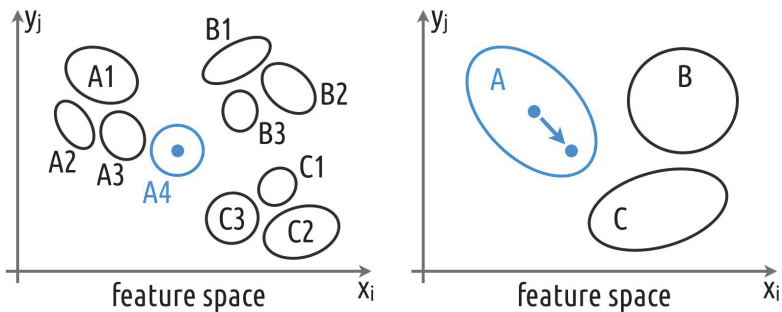
$$(\hat{\lambda}, \hat{q}) = \arg \max_{(\lambda, q)} \{ p(\lambda, q|\mathbf{x}) \} \quad (18)$$

using the Viterbi algorithm. Using Bayes' theorem, we obtain  $p(\lambda, q|\mathbf{x}) = p(\mathbf{x}|\lambda, q) \cdot p(\lambda, q)$  and see that this classification stage allows the introduction of external knowledges, via the use of prior probabilities on class sequences, which could be used to constrain the mapping (in speech recognition, this term would be related to the language model).

2.  $p(\mathbf{y}|\lambda, q, \mathbf{x})$  which is related to the *synthesis stage* and could be maximised similarly to as the GMM-based mapping technique, using Eq. 14. This is similar to the trajectory GMM technique, but here a continuity constrain is imposed on the estimated feature trajectories. The corresponding HMM is thus often referred to as a *trajectory HMM*. The modification proposed in the short-term MLPG can also be applied at this level.

## 2.5 Classification vs. Mapping: A Modelling Trade-Off

The integration of HMM-based classification, mapping and generation within the same framework gives us a chance to discuss about a specific modelling trade-off. Indeed from the same feature space (input and target features), various modelling strategies can be applied. On the one hand, we can create a large amount of small clusters. In that case, the labelling is very precise, the classification task needs to be very discriminative and the parameter generation has a very narrow area from which to get its target values, limiting the influence of the mapping. On the other hand, we can create a small amount of large clusters. In that case, the labelling is looser, discrimination between classes is easier and the parameter generation has a wide area from which to get its target values, hence much more relying on the mapping to browse the subspace. A summary of this trade-off is given in Fig. 3. The two approaches will lead to very different kinds of applications.



**Fig. 3.** Explanation of the modelling trade-off between classification, mapping and generation. We can choose between many small clusters where mapping is limited or few big clusters where mapping is primordial to browse the subspace.

## 3 Description of Data Types Used for Modelling

One main objective of this project was to design and develop a framework where GMMs and HMMs can be applied to a very great variety of data types. In this Section, we give a more exhaustive description of the data types that we have addressed and the databases that we have used. It gives a solid ground for understanding the feature spaces that we are manipulating.

### 3.1 Speech

The voice models that we used for speech are either identical or similar to the ones found in the HTS software [13]. Indeed our work with the speech databases has mainly used standard HTS training but it has also aimed at retraining some



voices, such as the ones used by the incremental speech synthesis application detailed in Section 4.1. For practical reasons, during development we generally work with the standard voice model from HTS demo scripts which is SLT, an American English female speaker from the CMU ARCTIC database [24], sampled at 48 kHz, 16-bit PCM. The training parameters for these models are the default ones in HTS for a sampling rate of 48 kHz. The input audio signals are split into 1200-sample long frames with a frame shift of 240 samples. For each frame, a parametric representation is extracted. The vocal tract is modelled by a 35-order MLSA filter [25] whose parameters are  $\alpha = 0.55$  and  $\gamma = 0$ , whereas the glottal source is described by a voiced/unvoiced decision and a pitch value, if any. On a higher level, each file from the training dataset has a phonetic transcription, from which phoneme duration models are trained, as well as additional information about syllables, words, sentences, accentuations, etc. [26]. For each parameter, this contextual information is used to train binary decision trees whose leaves are Gaussian models of the parameter.

### 3.2 Singing

Many similarities exist between speech and singing voice, though the differences are significant, especially for analysis and modelling. Some of the more prominent phenomena specific (though not necessarily exclusive) to Western classical singing<sup>2</sup> in contrast to regular speech are the higher voiced/unvoiced ratio, vibrato, higher range of loudness and pitch, singer’s formant and modification of vowels at high pitches [27]. These and other differences between speech and singing voice lead to some challenges in analysis modelling of singing.

The assumption of a decoupled source-filter is reasonably accurate especially for speech, but source-filter interactions are known to occur in various situations. For example, coupling effects increase by high pitch or high-impedance vowels like [u] [28]. Another challenge is that speech analysis algorithms such as pitch or *Glottal Closure Instant (GCI)* estimators often do not work as intended and this results in loss of performance [29,30]. Finally prosody models based on linguistic and phonetic information are almost never applicable due to the nature of singing. Instead different prosody models that take musical gestures and score information into account may be necessary.

In order to capture these singing-specific stylistics, we designed and recorded a new singing database for our purposes. The main challenge was to define the stylistic space in a meaningful manner. After consulting different systems of categorising singing voice, we decided that our approach is to set a *cognitive target* for the singer. Consequently we only specified more abstract stylistic information in the database design and aimed to capture the singing phenomena as they occur naturally. We contracted a professional female singer, EB, and the recordings were done in a professional recording studio. Seven pieces were recorded with multiple renditions in three different styles: *neutral*, *classical* and

---

<sup>2</sup> In our work we constrain our domain to Western classical singing, because it is well-formalised and a sizeable amount of previous technical work exists.

*belting*. Since naturalness was high priority, the pieces were chosen from EB’s repertoire of jazz, classical, folk and contemporary music, which she was comfortable and confident in performing. A limited amount of pure-vowel recordings were also done, consisting of some selected pieces and specific pitch gestures such as flat and constant ascent/descent. Contemporaneous electroglottography (EGG) recordings were also made, in order to establish ground truth for pitch and enable GCI-synchronous analysis methods. The resulting database is more than 70 minutes long, containing a rich variety of singing samples.

### 3.3 Audio-Visual Laughter

For addressing laughter modelling, the AV-LASYN database was used. The AV-LASYN database is a synchronous audio-visual laughter database built for the purpose of audio-visual laughter synthesis. The next paragraphs give an overview of the recording pipeline. Audio data was recorded at high sampling rate (96kHz) for eventual study of the impact of sampling rate on audio synthesis results. However, since it is a higher sampling rate than what common applications and research such as the present work need, we downsampled to 44.1kHz. Visual laughter was recorded using a marker-based motion capture system commercially available and known as OptiTrack. A set of 6 infrared cameras were used to track at 100 fps the motion of 37 markers glued on the subject. A seventh camera was used to record a grayscale video synchronised with all others. Among the 37 tracked markers are 4 markers placed on a headband. These helped to extract head motion from face deformation and make both available independently. After this separation process, we end up with 3 values for each facial marker ( $xyz$  coordinates) which corresponds to 99 values at each frame as well as 6 values at each frame that represent head motion ( $xyz$  coordinates and rotations around the same axes). This makes a 105-dimensional vector to represent overall face motion for a given frame. Neutral pose and the whole set of data corresponding to visual motion have been saved in the Biovision Hierarchy (BVH) format. The final corpus is composed of 251 segmented laughs. This corresponds roughly to 48 minutes of visual laughter and 13 minutes of audible laughter. For each laugh, the corpus contains: an audio file [44.1kHz, 16 bits], a BVH motion file that can be loaded in common 3D software (with the neutral pose, 6 channels for head motion, 3 channels for each of 33 facial markers), a binary motion file containing the same data as in the BVH to make it easier to load, a HTK label file containing phonetic transcriptions and temporal borders for each laughter phone.

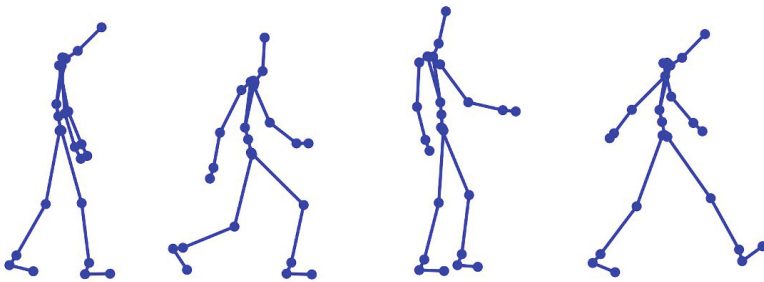
### 3.4 Audio-Visual Affective Speech

Experiments were conducted on the BIWI 3D Audiovisual Corpus of Affective Communication [40] comprising a total of 1109 sentences (4.67 seconds long on average) uttered by 14 native English speakers (6 males and 8 females). The dense dynamic face scans were acquired at 25 frames per second by a realtime 3D scanner and the voice signal was captured by a professional microphone at a sampling rate of 16kHz. For the voice signals, fundamental frequency, signal

intensity and segment duration are also provided. Along with the detailed 3D geometry and texture of the performances, sequences of 3D meshes are provided, with full spatial and temporal matching across all sequences and speakers. For each speaker 80 utterances are recorded, half in a personal speaking style and half in an “emotional” manner, as they are asked to imitate an original version of the performance.

### 3.5 Stylistic Gait

In this work, we also used the Mockey database [31] as our stylistic gait motion capture database. This database was recorded using a commercial inertial motion capture suit called IGS-190, from Animazoo [32]. This motion capture suit contains 18 inertial sensors, which record the angles between “body segments” corresponding to a simplified human skeleton representation. The output of the motion capture suit are these angles, expressed in the Euler angle parameterisation, and the calculated 3D position of the root (hips), which is computed given the angles and the lengths of the leg segments. In the database, the walk of a professional actor impersonating different expressive styles was recorded. The eleven styles represented in the database were arbitrarily chosen for their recognisable influence on walk motions. These styles are the following: proud, decided, sad, cat-walk, drunk, cool, afraid, tiptoeing, heavy, in a hurry, manly. Each walk style is represented by a different number of steps in the database, ranging from 33 to 80 steps. Fig 4 gives an overview of these walking styles.



**Fig. 4.** Generic skeleton frames extracted from the Mockey database [31] and corresponding to different styles of gait, here: sad, afraid, drunk and decided

The Mockey mocap sessions are recorded in the Biovision Hierarchy (BVH) format. The skeleton from the Animazoo software is defined by 20 body segments, and each data frame hence contains 66 values. Three of the segments are in fact only added to make the simplified skeleton look closer to a real skeleton but have no degree of freedom in the motion capture. There are hence finally 57 values to analyse and model in our data, among which 54 joint angle values and 3 values for the skeleton root Cartesian coordinates. Since the Cartesian

coordinates of the root are calculated a posteriori from the joint angles, we only took into account the 54 joint angle values in our models. Furthermore, since the Euler angle representation is seldom optimal, we converted it to the exponential map angle parameterisation, which is locally linear and where singularities can be avoided. The Mockey database has been recorded at a rate of 30 frames per second. The walk sequences have been automatically segmented into left and right steps, based on the evolution of the hip joints angles.

## 4 Realtime and Reactive HMM-Based Generation

The first step in creating our new framework was to validate the adaptation of the realtime HMM-based parameter generation – such as described in Section 2 and implemented in MAGE – to the new data types described in Section 3. It brought us to implement a series of prototypes that are described in this Section.

### 4.1 Incremental Speech Synthesis

In several applications of Text-To-Speech (TTS) it is desirable for the output to be created incrementally. Such applications include reactive dialogue systems and speech devices for disabled people. However current TTS systems rely on full pre-specified utterances provided before the synthesis process begins, severely limiting the reactivity and realtime use of speech synthesis. While MAGE is capable of realtime synthesis, it is reliant on linguistic context labels which are computed offline prior to synthesis. This means the utterance to be synthesised is fixed and cannot be changed at run-time except to other pre-computed context labels. There is, however, nothing to prevent MAGE from synthesising from an incrementally created set of labels. Hence a new realtime linguistic front-end was created in order to allow for continuous incremental creation of the linguistic context labels for synthesis. A new front-end was chosen as current front-ends are simply nowhere near fast enough for realtime analysis – e.g. Festival takes over 1000ms to process an utterance (even a single word) and MARY-tts slightly above 200ms [33] – and they assume the full utterance is present at analysis time.

**Linguistic Analysis.** The standard full-context labels used by the HTS engine includes a large amount of varying contexts used in the decision tree context clustering process. Many of these do not lend themselves to incremental processing as they rely on the presence of the full utterance. Therefore a reduced context set was decided upon based on the standard HTS set [34]. Any contexts related to the full utterance or phrases were removed as these are not available. Contexts regarding future syllables are included ‘if available’ and the system requires two future phones. The decision to retain two future phones, making the system in effect lagging two phones behind realtime, was made as informally listening to the speech when no future phones were included resulted in a significant degradation of the speech intelligibility. The resulting context labels included 23 contexts down from 53. Informally no noticeable drop in quality was perceived

on a voice re-trained on the reduced context set compared to a voice trained on the standard HTS contexts. The system works with word-sized chunks, such that every time a user inputs a complete word the system will provide the labels necessary to synthesise the word, with a minimum of two phonemes. The words are looked up in the CMUDict 0.4 dictionary from which stress patterns, syllables and phones are retrieved and the labels created. No letter-to-sound rules are included. If a word is not be in the dictionary, a filled pause is introduced instead (“uh”).

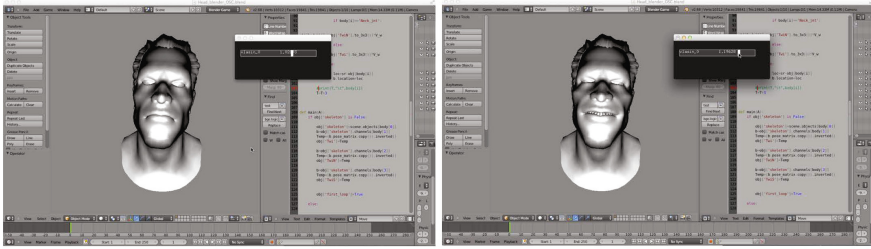
**Typing Interface.** A simple typing interface was implemented which allows a user to type in the string to be synthesised. It is however incredibly difficult to type as fast as the synthesis speed, so synthesis was slowed by a factor of 2.5 to allow a skilled typer to type fast enough. To further enhance the ability of the typist to type quickly, simple word prediction was added allowing the user to press the F1 to F5-keys to instantly insert a word predicted by the system.

## 4.2 Reactive Control of Audio and Visual Laughter

In this part we explain how the modelling of the laughter has been approached with HMMs, both for the sound and the face motion. Then we give a first description on how we turned these two processes into an interactive application.

**Reactive Acoustic Laughter Synthesis.** HMM-based acoustic laughter synthesis is a problem that has been addressed recently [35,36]. The same pipeline has been applied in this work to train an acoustic model of laughter using the AV-LASYN database described in Section 3.3. We have extracted 35 Melcepstral coefficients and log F0 as features to represent the acoustic signal. We used STRAIGHT [37] for this extraction process. Then, five-state, left-to-right HMMs were trained to model a set of laughter phones (see [35] for more information). From the synthesised F0, an excitation signal was derived and modified by DSM [38]. Finally synthesised MFCC trajectories and modified excitation signal were used to feed a source-filter model and generate the corresponding waveforms. With the acoustic models obtained as explained above, we were able to integrate acoustic laughter into MAGE. Although there is still room for improvements, we have shown that reactive acoustic laughter synthesis is feasible through MAGE. Further investigation is needed to have a better understanding of the behaviour of MAGE for the synthesis of the specific signal of laughter, and this work will serve as a basis for further studies.

**Visual Laughter Synthesis.** A similar process has been applied to visual data contained in the AV-LASYN database. As a reminder of Section 3.3, the visual data consists of facial points trajectories. Thirty-three markers on the face represented at each frame by 3 spatial coordinates are used. Six other values represent the head motion. The features that we used to train the HMMs were obtained



**Fig. 5.** Illustration of the reactive control of laughter intensity by having a MAGE application to send trajectories to the 3D face model in Blender through OSC

as follows. First we subtracted the neutral face from each frame so that the data represents only the facial deformation. Then a PCA analysis was performed and showed that 97% of the variability in the data is contained in the 4 first principal components. We hence decided to reduce the dimensionality of the data to 4. However the PCA analysis did not include the 3 values representing the head rotation for a matter of consistency of the data on which PCA was applied. We thus end with a 7 dimensional feature space to represent visual data, instead of the 105 dimensional original space. In order to train the HMMs, annotations are needed. First the acoustic annotations provided in the AV-LASYN database were used but we quickly came up with the conclusion that annotations based on audio are not suitable for visual modelling. We then tried to annotate manually a subset of the database based on the visual behaviour. Three classes were used: *laugh*, *smile* and *neutral*. The results of the training based on these new annotations gave much better results. Since annotating manually is a highly time consuming task, we have tried to do it automatically, using clustering techniques. Gaussian Mixture Models were used for this purpose. A GMM was fitted to the whole visual data and each frame was classified among 3 clusters based on this fitting. From this classification, we derived visual annotations to be used in the HMM training. The resulting trajectories appeared to be plausible facial laugh motion.

**Reactive Visual Laughter Synthesis.** As we did for audio, we then tried to integrate these models into MAGE to be able to synthesise facial motion reactively. Therefore we had to add a module to MAGE so as to be able to project back the synthesised trajectories into the high dimensional original space. After this projection, the data is available in a format which may be retargeted on a 3D face model. This was done by using Blender in which we loaded an already-rigged 3D face model. Data is sent trough OSC from MAGE to Blender where it is read and applied to the 3D face with a python script. As a proof of concept, we decided to synthesise a succession of neutral and laughing faces in a loop. We also added a control parameter that allows to change the intensity of the visual laughter in realtime. This control parameter amplifies or attenuates the

generated trajectory dynamics. An illustration of the reactive visual laughter in Blender is given in Fig. 5.

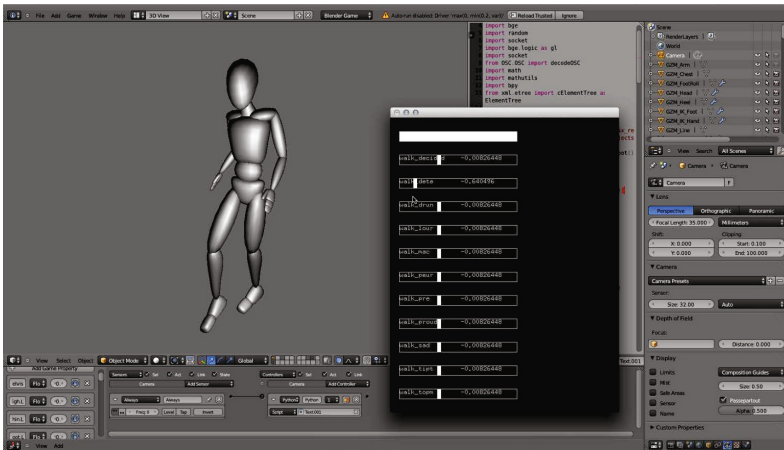
### 4.3 Reactive Exploration of a Stylistic Gait Space

Motion style is something difficult to capture since it is hardly describable. Our human expertise enables us to decode effortlessly the emotion, quality or style conveyed in otherwise functional motions. However it is almost impossible to formally describe the alterations which, once applied to the functional basic motion, give it its specific style. Furthermore, making the distinction between the variability of human motion execution and the style of the motion itself is an additional difficulty when aiming at modelling the style of a motion. Indeed, when performing twice the same motion with the same style, the execution of the movement will always slightly vary. In this work we implemented a framework for stylistic exploration of motion, using the expressive walk case study as a proof-of-concept. Our approach is to foster the generative exploration of styles, from statistical models, as a way of highlighting their implicit properties.

**Stylistic Gait Modelling and Synthesis.** The statistical nature of HMMs enables them to take into account the intrinsic variability of execution of human motion. Both the duration variation and the execution variation are modeled, and a HMM trained on stylistic mocap data becomes a summary of that particular style. Using the Mockey database as training data, the walk was modelled by one five-states left-to-right HMM per step (left and right), following the approach presented in [14]. In a first phase, a global model was trained using all the database. In a second phase, an adaptive training was conducted in order to adapt the generic walk model to each one of the eleven styles present in the database, giving a total of eleven style-specific walk models and one neutral global model. Such an approach corresponds to the left side of Fig. 3, as described in Section 2. In these models, a diagonal covariance matrix is used when modelling the pdfs of the observations, hence not taking into account the interdependency existing between the different body joints motions. Each one of these models can be used to synthesise new walk sequences of any chosen length, and the generated walk sequences will display the style of the models from which they have been generated.

**Continuous Stylistic Gait Synthesis.** However in addition to style and motion variability, the alterations of the functional motions not only convey the style of the motion, but also the intensity of expression of that specific style. Since that intensity can vary continuously, it is impossible to capture the whole range of intensity during motion capture sessions, even for one single style. With our twelve gait models, we are able to generate walk sequences which display the same styles as the ones present in the training database, plus one “neutral” style trained on all the styles. However since all of our models present the same structure, as they have all been adapted from the same generic model, we can

take advantage of this alignment in order to produce new walk styles which have not been recorded. The model parameters space (mean and variances of output and duration pdfs) is considered as a continuous stylistic space, where the values corresponding to each recorded style can be viewed as landmarks which tag the space. Through interpolation and extrapolation between these landmarks, we are able to obtain new walk style models. The intensity of each style can be controlled by interpolating between the original style and the neutral style parameters, also enabling the production of exaggerated and opposite styles. Completely new walk styles can also be built by combining any of the existing styles, enabling the free browsing of the complete stylistic walk space. This approach has been validated in [39]. However in this work both the control of the style and the walk synthesis were implemented as offline processes, preventing the desired free interactive user exploration.



**Fig. 6.** Illustration of the application for gait style exploration: the MAGE application sends trajectories corresponding to interpolated gait models to Blender through OSC, where the 3D character is animated. The MAGE interface gives one slider per style.

**Reactive Stylistic Gait Synthesis.** In the current work, we implemented a reactive gait style exploration application, enabling the user to reactively control the style of the synthesised walk thanks to MAGE, and to visualise the resulting motion sequence in realtime. In this application, the user browses this stylistic space in realtime, through a set of sliders controlling the influence of each original style, as illustrated in Fig. 6. These stylistic weights are sent to MAGE, which synthesises an infinite walk sequence (a loop of left and right steps), and the walk model is adapted in realtime with the weights corresponding to the sliders, hence modifying the style of the synthesised walk. The synthesised walk trajectories are sent to Blender through OSC, where it is displayed in realtime on a virtual



3D character. The user is hence given interactive control of the walk of a virtual character, as he manipulates sliders to control the style, and can see the influence of these stylistic modifications on the walk of the Blender virtual character. This proof of concept application opens the doors to many possibilities as the size of motion capture databases nowadays explodes and more and more applications seek new possibilities for exploring motion style or compare motions.

## 5 Realtime HMM-Based Continuous Mapping

The second step in creating our new framework was to validate some mapping strategies within the HMM-based approach – such as described in Section 2. Therefore we have developed a few use case prototypes where the user control was captured and decoded gestures. This Section explains these applications.

### 5.1 Audio-Visual Face Retargeting

Speaker identity conversion refers to the challenging problem of converting multimodal features between different speakers so that the converted performance of a source speaker can be perceived as belonging to the target speaker. We addressed the problem of speaker conversion using audio and 3D visual information. The speech signal and the 3D scans of a source speaker for a certain utterance will be modified to sound and look as if uttered by a target speaker. The speaker-specific features are mapped between a source and a target speaker using GMMs, as described in Section 2.

In the offline version, the 3D BIWI dataset [40] is used to train the GMM model between any two speakers. The training is done on 40 utterances performed in a neutral manner by both speakers. Spectral features are extracted at a segmental level using the STRAIGHT vocoder [37] which decomposes speech into a spectral envelope without periodic interferences, F0 and relative voice aperiodicity. From the spectral envelope, we use the 1st through 24th Mel-cepstral coefficients, a widely made choice in voice conversion and voice synthesis/analysis systems. The speaker-specific facial articulation features are captured from a dense mesh of 3D data. From the dataset, 7 speech and expressive movement components are extracted following a guided PCA method [41]. As the mouth opening and closing movements have a large influence on face shape, the first jaw component is used as a first predictor, iterative PCA is performed on residual lips values and the next 3 lips components are obtained. The second jaw component is used as the 5th predictor and the last two parameters are extracted as expressive components and represent the zygotic and eyebrow muscle movements. These features are computed at the original video frame rate and are later oversampled to match the audio frame rate. Both visual and spectral features are concatenated with their first derivatives in order to be used for the MLE-based mapping approach.

With the purpose of creating an interactive scene in which virtual actors are able to interact in realtime as guided by a director, we are looking into the possibility of a realtime conversion framework. For this framework, a new system

setup is used, involving a Primesense Carmine 1.09 camera to capture close-range face expressions and a microphone for audio signal acquisition. Also new approaches are needed for extracting relevant audio-visual features. Therefore the Faceshift software [43] is used to generate a realtime 3D face mesh while a speaker performs in front of the camera. For each frame, 48 parameters that control different face movements (jaw opening, eye squint, puff, sneer etc) are also generated. They are called *blendshapes* and can be used with the associated mesh of the user or retargeted to an existing mesh. In the case of audio features, a realtime version of the SPTK tool and MLSA filter are implemented for extracting MFCC coefficients and audio synthesis.

The GMM models are trained offline on a database composed of recorded audio signals of two speakers and the associated blendshapes generated from Faceshift. The converted blendshapes can also be sent to Blender to create a realtime face animation using the 3D mesh of the target speaker. The communication between the different softwares in realtime is done in Max. Like the SPTK and MLSA realtime tools, GMMmap is a module for gaussian mixture model regression using the MMSE method implemented in Max. It uses the models that were trained offline and saved in a suitable format and the audio-visual features that are extracted in realtime to estimate the converted features.

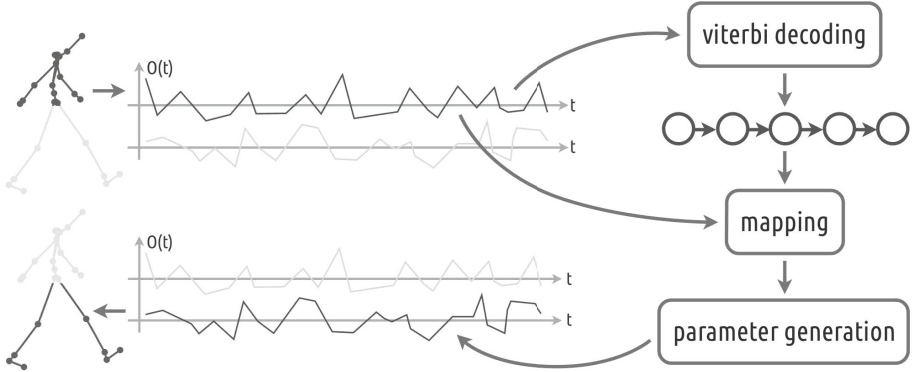
## 5.2 Realtime Stylistic Full-Body Gait Reconstruction

In our application for exploring the stylistic gait space described in Section 4.3, the ongoing motion is created by the linear combination of the twelve distinct stylistic walk models, according to the weights given to each style on the GUI. Such an approach requires that the stylistic space is explicitly tagged according to the names used in the training database and proposes a vision of the continuum between styles based on linear interpolation.

However considering that the various styles can be named and interpolated within the feature space is a strong design decision. Many use cases might benefit from more implicit approaches towards stylistic exploration. Particularly we wanted to give the user the ability explore the stylistic space through HMM-based mapping between his/her input gestures and the corresponding output. With this idea, we refer to the right side of Fig. 3 where mapping plays the important role in browsing the feature space, as described in Section 2.

In order to validate that the inherent style of a motion can be determined from a subset of its dimensions and remapped in realtime on the remaining dimensions, we have built a prototype that will reconstruct the gait (step sequence plus style) from the upper to the lower body. It means that each 54-channel (18 nodes, each with 3 angles) feature vector from the Mockey database is actually split into inputs and outputs. We consider that the 36 channels corresponding to the upper body (from head to hips) are inputs. The other 18 channels corresponding to leg joints are considered as outputs. They will be animated in realtime by the system. The whole process is illustrated in Fig. 7.

To achieve the regression between upper and lower body dimensions, we implemented a HMM-based mapping as explained in Section 2. The sequence of



**Fig. 7.** Illustration of the overall process used in the gait reconstruction example: continuous inputs are decoded with a realtime Viterbi algorithm. This decoding generates an ongoing state sequence that is used for synthesis of the outputs. Before pdfs are used for synthesis, means are modified by a mapping function based on covariance.

inputs  $\mathbf{x}$  are the channels of the upper body and the sequence of target feature vectors  $\hat{\mathbf{y}}$  to be estimated are the channels of the lower body. The gait models used are trained on all the styles with full covariance matrices in the pdf representation of the observations. We have a HMM for the right step and a HMM for the left step. Each HMM owns four states.

The first stage in this process is the decoding of the input sequences. The implemented solution for the decoding uses the HTK software toolkit [45]. A realtime data stream is simulated by sending the input data with a specified frame rate to the HREC algorithm, a HTK module that applies the Viterbi algorithm. We added, in the pdfs of the observations, a mask vector to inhibit the channels corresponding to the outputs of the mapping process. This decoding stage provides the most likely HMM  $\hat{\lambda}$  that is being played by the streamed data and the current state for this model. To ensure that this stage works in realtime, we extract the partial results given by the Viterbi algorithm without being sure to have the realisation of the complete states sequence for a given model. Moreover, for a given frame  $\mathbf{x}_t$ , only its past  $[\mathbf{x}_1, \dots, \mathbf{x}_{t-1}, \mathbf{x}_t]$  is taken into account to establish the current state  $\mathbf{q}_t$ . It appears that it could be more accurate to compute this state by introducing a delay in order to get some information about the future of each sample to choose the best states sequence.

Once the decoded state is available, it can be used to query the HMM database of the upper body dimensions so as to build the state sequence for the synthesis stage. Before the stack of pdfs is accumulated for synthesis, the means of each state are extracted and corrected according to the mapping function described in Section 2. This process tends to influence the means so as to move within the model and react to the covariance information which is expressed between the input and output dimensions. As a result, the statistical properties of the state sequence get modified. When this modified state sequence enters the synthesis

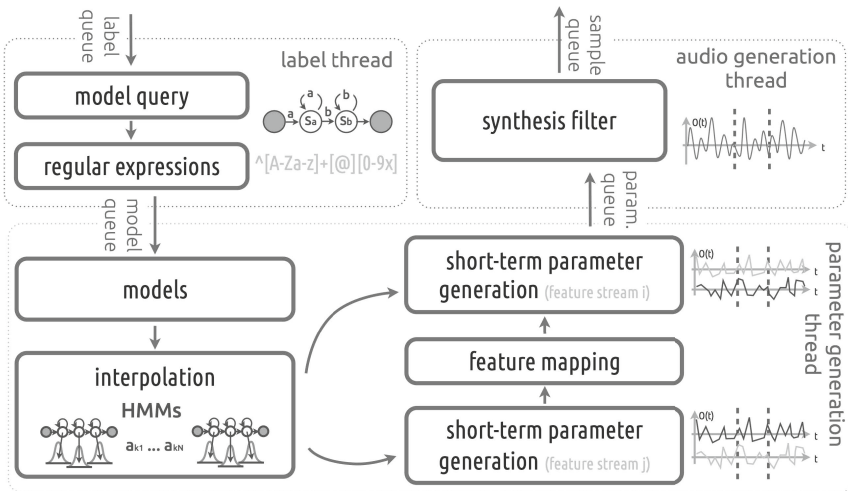
stage, it reflects the stylistic influence of the inputs on the outputs. It means that the style of the upper body transfers to lower body trajectories.

## 6 Architecture and Software

Based on the reactive properties of HMM-based speech synthesis framework, as described in [26], we built a new speech synthesis library, called MAGE [23]. MAGE is based on the HMM-based parametric speech synthesis system (HTS), which it extends in order to support realtime architecture and multithreaded control. As it is based on HTS, it inherits its features, advantages and drawbacks [26]. The contribution of MAGE is that it opens the enclosed processing loop of the conventional system and allows reactive user control over all the production levels. Moreover, it provides a simple C++ API, allowing reactive HMM-based speech synthesis to be easily integrated into realtime frameworks [46,47].

### 6.1 Threaded Architecture of MAGE

One important feature of MAGE is that it uses multiple threads, and each thread can be affected by the user which allows accurate and precise control over the different production levels of the artificial speech. As illustrated in Fig. 8, MAGE integrates three main threads: the *label thread*, the *parameter generation thread* and the *audio generation thread*. Four queues are shared between threads: the *label queue*, the *model queue*, the *parameter queue* and the *sample queue*.



**Fig. 8.** MAGE: reactive parameter generation using multiple threads and shared queues

The *label thread* controls the phonetic labels, by pulling the targeted phonetic labels from the *label queue* and pushing the corresponding models into the *model*

*queue*. It is responsible for the contextual control of the system. The *parameter generation thread* reads from the *model queue* a model that corresponds to one phonetic label at a time. For that single label / model the speech parameters are generated (static and dynamic features), which are locally-maximised using only the current phonetic label / model (and if available, the two previous labels). In other words, for every single input phonetic label, the feature vectors are estimated by taking into account the HMMs of that specific label. The generated speech parameters are stored in the *parameter queue*. Finally, the *audio generation thread* generates the actual speech samples corresponding to the inputted phonetic label and store them in the *sample queue* so that the system's *audio thread* will access them and deliver them to the user. Further details of the MAGE reactive parameter estimation can be found in [49].

## 6.2 Reactive Controls

Accessing and controlling every thread has a different impact over the synthesised speech, as illustrated in Fig. 9. The *label thread* can provide contextual phoneme control. Indeed, the context of the targeted output can be easily manipulated in realtime by simply controlling which of the available phonemes for processing will be inputted into the system and in which sequence.

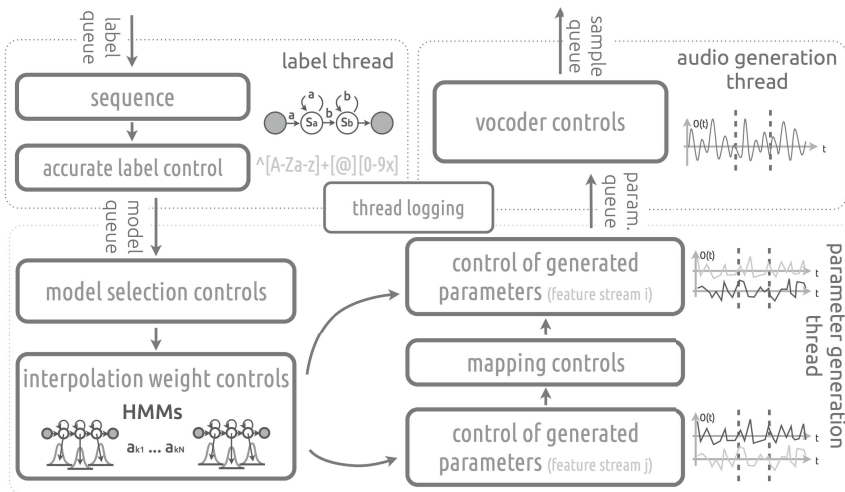
The *parameter generation thread* can reactively modify the way the available models are used for the parameter sequence generation [49]. The user can reactively alternate between the available models, or interpolate between models with different interpolation weights among the various feature streams. It is also possible to manipulate the mapping between different feature streams, i.e. how a given stream influences another [17].

Finally the *audio generation thread* manipulates reactively the vocoding of every sample, resulting in prosody and voice quality controls. The delay in applying any received user control varies between a single speech sample and a phonetic label depending on the thread that is being accessed. For every thread it is also possible to enable the logging functionality, described hereafter, to store the applied user controls as well as its generated output.

## 6.3 Reactive Control through Regular Expressions

One new feature added in MAGE is the support of regular expressions. As explained in [26], in order to describe a phoneme, additional linguistic specifications have to be taken into account. Therefore, every phonetic label in addition to phoneme information, uses various linguistic contexts such as lexical stress, pitch accent, tone, and part-of-speech information for the context-dependent modelling. An example of the format of the phonetic labels can be found in [13].

Until now, it was possible to control the sequence of the labels inputted to MAGE be synthesised. However, there is need for more accurate and specific control over the targeted context. In order to achieve that we use regular expressions that describe the phonetic labels. The integration of regular expressions



**Fig. 9.** MAGE: reactive user controls over the multiple threads

“allows the user to query” every imputed label and accordingly to apply certain controls, on every production level of the artificial speech.

For example, when it comes to the contextual control, that occurs in the *label thread*, if the current phoneme is “*v*” the synthesis of that phoneme can be skipped. Another example, while controlling the models themselves through the regular expressions, a control that occurs respectively at the *parameter generation thread*, if the next phoneme is “*ə*” we want to interpolate speaking style *i* with speaking style *j* using interpolation weight vector  $\mathbf{y}$ . Finally, while controlling the actual generated speech samples, accessing the *audio generation thread*, if the phoneme is stressed then the pitch can be shifted, etc.

## 6.4 Reactive Mapping Control

In previous versions of MAGE, the granularity of the controls that the user can access for the parameter generation stopped at the model level. Indeed, through the API, a user could only push left-to-right HMMs into the model queue and, from there, compute the duration of each state and the sequence of corresponding observations. This constrains the use case into a left-to-right pattern that does not allow integration of the mapping control detailed in Section 5. Therefore, we added a state queue into MAGE as an alternative to the model queue.

While the model queue is usually filled with sequences of states corresponding to models selected to match the labels in the label queue, the state queue is fed directly with one state at a time. Each state corresponds to one frame of observations and, as such, has a duration of one. If the system must remain in a state for  $N$  frames, that state is simply pushed in the queue  $N$  times. This enables arbitrary patterns and number of states for the HMMs and thus overcomes the limiting effect of the model queue.

As for the short-term computation of observation frames from the state queue, it is achieved almost as for the model queue. The most significant difference is that one can set  $M$ , the number of frames to be computed whereas in the model queue the frames are computed for one complete model at a time, and the number of frames generated is equal to the duration of that model. The context for the short-term parameter computation is set in the same fashion as for the model queue, except that the user sets a number of states to be considered before and after instead of a number of models. This notably allows to always use a constant amount of contextual information, for instance 17 states in the past and 3 states in the future of the  $M$  states that correspond to the  $M$  frames to be computed. This contrasts with the model queue for which the amount of contextual information is the sum of the durations of each model used as context and thus can change at every step. Using the state queue with  $M = 1$ , one can even make the computation for one state at a time. In other words, one can generate one frame at a time, while still using surrounding states as the context for the short-term parameter computation.

## 6.5 Logging of User Actions and Generated Output

One of the major complications when working with reactive applications is that of detecting and explaining unexpected situations. Indeed, every action from the user can cause an instant, or not so instant, reaction from the system. Depending on many factors (which MAGE's thread it is applied to, OS process scheduling policy, etc.), both the reaction and its time delay after the action occurs can be different, even if the user reproduces the same set of interactions. Therefore, when something unusual happens it can be very difficult to, first, realise it and then reproduce it to eventually pinpoint the cause of the event. It could simply be a normal, albeit surprising, answer to a one in a million combination of user commands but it might as well be a bug in the application or in the core MAGE library. Added to this is the problem of detecting exactly when it happened.

In order to make it easier to solve these issues, we introduced a simple logging system in MAGE, as illustrated in Fig. 9. If enabled, it records the sequences of controls sent to MAGE by the user such as the labels, pitch,  $\alpha$ , interpolation weights, etc. Each of these values is recorded with a timestamp corresponding to the index of the first sample to which it is actually applied inside of MAGE. Besides, the logging system also saves the evolution of the inner state of MAGE. This is currently limited to the content of the frame and sample queues but could easily be extended to the model and state ones.

## 7 Conclusions

In this project we have gathered different approaches and backgrounds, with the common aspect of being interested in applied statistical modelling, and we have created a unified framework. This framework is based on trajectory GMMs and HMMs and use a newly-created gesture recognition tool and a new version

of MAGE in order to enable the development of mapping strategies. The idea of HMM-based mapping has been formalised and generalised to any kinds of input and target stream of features. Such a reflexion had a significant impact in how we were envisioning the use of generative models in this work. Therefore we have been able to create a first set of new prototype applications to assess our approach of parameter generation. Indeed we have created an incremental speech synthesiser, generating speech audio right when the user is typing with a limited delay. The current incremental synthesis system allows for a simple analysis relying on a lexical look-up to provide simple lexical analysis and word prediction. This sufficiently demonstrates the possibility of realtime incremental TTS. Many possible future directions can be perceived, such as the implementation of better user interfaces for faster input to the system, the utilisation of MAGE’s realtime speech modifications capabilities (e.g. to adjust synthesis speed to user input speed) and the prediction of future phones (and other contexts) to allow the system to be truly realtime without a significant loss of synthesis quality. Also in the parameter generation improvements, we have created realtime exploration of mocap-based trajectories. This idea has been applied to face and body. For the face, it gave the first realtime audio and visual laughter prototype. For the body, we created a new application for exploring the stylistics of gait by blending together various identified one-style models and enable inhibition and exaggeration of those styles. The second step in our work has been to design and assess more implicit statistical mapping applications, where the input is a natural gesture. There we have developed a audio-visual speaker retargeting prototype, where the expressive multimodal speech gestures of a given speaker are remapped on another one in realtime. Also we have created the first gait reconstruction application, where the upper body gait (balancing arms and torso) triggers the parameter generation corresponding to the lower body motion (legs) in realtime. This prototype demonstrates that that HMM-based recognition of stylistic data, its mapping and the corresponding parameter generation can be achieved in a realtime scenario. Finally most of these developments have helped the MAGE software to head towards its third major release.

**Acknowledgement.** N. d’Alessandro is funded by a regional fund called Région Wallonne FIRST Spin-Off. J. Tilmanne and T. Ravet are supported by the European Union, 7th Framework Programme (FP7-ICT-2011-9), under grant agreement n° 600676 (i-Treasures project). M. Astrinaki and O. Babacan are supported by a PhD grant funded by UMONS and Acapela Group. H. Çakmak receives a PhD grant from the Fonds de la Recherche pour l’Industrie et l’Agriculture (FRIA), Belgium. The work of Adela Barbulescu has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025).

## References

1. Mori, M.: The Uncanny Valley. *Energy* 7(4), 33–35 (1970)
2. Mori, M.: The Uncanny Valley (K. F. MacDorman & N. Kageki, Trans.). *IEEE Robotics & Automation Magazine* 19(2), 98–100 (2012)



3. Dutoit, T.: An Introduction to Text-To-Speech Synthesis. Kluwer Academic Publishers Inc. (1997)
4. Raux, A., Black, A.W.: A Unit Selection Approach to F0 Modelling and its Applications to Emphasis. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 700–705 (December 2003)
5. Lindemann, E.: Music Synthesis with Reconstructive Phrase Modelling. IEEE Signal Processing Magazine 24(2), 80–91 (2007)
6. Fechteler, P., Eisert, P., Rurainsky, J.: Fast and High Resolution 3D Face Scanning. In: IEEE International Conference on Image Processing, vol. 3, pp. 81–84 (2007)
7. Menache, A.: Understanding Motion Capture for Computer Animation and Video Games. Morgan Kaufman Publishers Inc. (2000)
8. d’Alessandro, N.: Realtime and Accurate Musical Control of Expression in Voice Synthesis. PhD defence at the University of Mons (November 2009)
9. Maestre, E., Blaauw, M., Bonada, J., Guaus, E., Perez, A.: Statistical Modelling of Bowing Control Applied to Violin Sound Synthesis. IEEE Transactions on Audio, Speech, and Language Processing 18(4), 855–871 (2010)
10. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), vol. 3, pp. 1315–1318 (2000)
11. Dutrevel, L., Meyer, A., Bouakaz, S.: Feature Points Based Facial Animation Retargeting. In: Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology, pp. 197–200 (2008)
12. Hunt, A., Wanderley, M., Paradis, M.: The Importance of Parameter Mapping in Electronic Instrument Design. Journal of New Music Research 32(4), 429–440 (2003)
13. Tokuda, K., Oura, K., Hashimoto, K., Shiota, S., Takaki, S., Zen, H., Yamagishi, J., Toda, T., Nose, T., Sako, S., Black, A.W.: HMM-based Speech Synthesis System (HTS), <http://hts.sp.nitech.ac.jp>
14. Tilmanne, J., Moinet, A., Dutoit, T.: Stylistic Gait Synthesis Based on Hidden Markov Models. Eurasip Journal on Advances in Signal Processing 2012(1,72) (2012)
15. Urbain, J., Cakmak, H., Dutoit, T.: Evaluation of HMM-Based Laughter Synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 7835–7839 (2013)
16. Astrinaki, M., d’Alessandro, N., Picart, B., Drugman, T., Dutoit, T.: Reactive and Continuous Control of HMM-based Speech Synthesis. In: IEEE Workshop on Spoken Language Technology (December 2012)
17. Astrinaki, M., Moinet, A., Yamagishi, J., Richmond, K., Ling, Z.-H., King, S., Dutoit, T.: MAGE - Reactive Articulatory Feature Control of HMM-Based Parametric Speech Synthesis. In: Proceedings of the 8th ISCA Speech Synthesis Workshop, SSW 8 (September 2013)
18. Hueber, T., Bailly, G., Denby, B.: Continuous Articulatory-to-Acoustic Mapping using Phone-Based Trajectory HMM for a Silent Speech Interface. In: Proceedings of Interspeech, ISCA (2012)
19. Kay, S.M.: Fundamentals of Statistical Signal Processing: Detection Theory, vol. 2. Prentice Hall PTR (1998)
20. Stylianou, Y., Cappé, O., Moulines, E.: Continuous Probabilistic Transform for Voice Conversion. IEEE Transactions on Speech and Audio Processing 6(12), 131–142 (1998)

21. Kain, A.B.: High Resolution Voice Transformation. PhD Thesis, Rockford College (2001)
22. Toda, T., Black, A.W., Tokuda, K.: Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Transactions on Audio, Speech, and Language Processing* 15(8), 2222–2235 (2007)
23. Astrinaki, M., Moinet, A., Wilfart, G., d'Alessandro, N., Dutoit, T.: MAGE Platform for Performative Speech Synthesis., <http://mage.numediart.org>
24. Kominek, J., Black, A.W.: CMU Arctic Databases for Speech Synthesis. Tech. Rep., Language Technologies Institute, School of Computer Science, Carnegie Mellon University (2003)
25. Imai, S., Sumita, K., Furuichi, C.: Mel Log Spectrum Approximation (MLSA) Filter for Speech Synthesis. *Electronics and Communications in Japan, Part I* 66(2), 10–18 (1983)
26. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech Synthesis Based on Hidden Markov Models. In: *Proceedings of IEEE*, vol. 101(5) (2013)
27. Sundberg, J.: The Science of Singing Voice. PhD Thesis, Illinois University Press (1987)
28. Titze, I.R.: Nonlinear Source-Filter Coupling in Phonation: Theory. *J. Acoust. Soc. Am.* 123, 2733–2749 (2008)
29. Babacan, O., Drugman, T., d'Alessandro, N., Henrich, N., Dutoit, T.: A Comparative Study of Pitch Extraction Algorithms on a Large Variety of Singing Sounds. In: *Proceedings of ICASSP* (2013)
30. Babacan, O., Drugman, T., d'Alessandro, N., Henrich, N., Dutoit, T.: A Quantitative Comparison of Glottal Closure Instant Estimation Algorithms on a Large Variety of Singing Sounds. In: *Proceedings of ICASSP* (2013)
31. Tilmanne, J., Ravet, T.: The Mockey Database, <http://tcts.fpms.ac.be/~tilmanne>
32. IGS-190, Animazoo website, <http://www.animazoo.com>
33. Baumann, T., Schlangen, D.: Recent Advances in Incremental Spoken Language Processing. In: *Interspeech 2013 Tutorial 1* (2013)
34. Oura, K.: An Example of Context-Dependent Label Format for HMM-Based Speech Synthesis in English. In: *HTS-demo\_CMU-ARCTIC-SLT* (2011), <http://hts.sp.nitech.ac.jp>
35. Urbain, J., Cakmak, H., Dutoit, T.: Evaluation of HMM-based Laughter Synthesis. In: *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP* (2013)
36. Urbain, J., Cakmak, H., Dutoit, T.: Automatic Phonetic Transcription of Laughter and its Application to Laughter Synthesis. In: *Proceedings of the 5th Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction* (2013)
37. Kawahara, H.: Straight, Exploitation of the Other Aspect of Vocoder: Perceptually Isomorphic Decomposition of Speech Sounds. *Acoustical Science and Technology* 27(6) (2006)
38. Drugman, T., Wilfart, G., Dutoit, T.: A Deterministic Plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis. In: *Proceedings of Interspeech* (2009)
39. Tilmanne, J., Dutoit, T.: Continuous Control of Style and Style Transitions through Linear Interpolation in Hidden Markov Model Based Walk Synthesis. In: Gavrilova, M.L., Tan, C.J.K. (eds.) *Transactions on Computational Science XVI. LNCS*, vol. 7380, pp. 34–54. Springer, Heidelberg (2012)

40. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: Acquisition of a 3D Audio-Visual Corpus of Affective Speech. *IEEE Transactions on Multimedia* 12(6), 591–598 (2010)
41. Bailly, G., Govokhina, O., Elisei, F., Breton, G.: Lip-Synching Using Speaker-Specific Articulation, Shape and Appearance Models. *EURASIP Journal on Audio, Speech, and Music Processing* 2009(5) (2009)
42. Barbulescu, A., Hueber, T., Bailly, G., Ronfard, R.: Audio-Visual Speaker Conversion Using Prosody Features. In: *International Conference on Auditory-Visual Speech Processing* (2013)
43. Faceshift, <http://faceshift.com>
44. Max Audio Software, <http://cycling74.com/products/max>
45. University of Cambridge, The Hidden Markov Model Toolkit (HTK), <http://htk.eng.cam.ac.uk>
46. Puckette, M.: Pure Data, <http://puredata.info>.
47. Lieberman, Z., Watson, T., Castro, A., et al.: openFrameworks, <http://www.openframeworks.cc>
48. Astrinaki, M., Moinet, A., d’Alessandro, N., Dutoit, T.: Pure Data External for Reactive HMM-based Speech and Singing Synthesis. In: *Proceedings of the 16th International Conference on Digital Audio Effects, DAFX 2013* (September 2013)
49. Astrinaki, M., d’Alessandro, N., Reboursiere, L., Moinet, A., Dutoit, T.: MAGE 2.0: New Features and its Application in the Development of A Talking Guitar. In: *Proceedings of the 13th International Conference on New Interfaces for Musical Expression, NIME 2013* (May 2013)

# Laugh When You're Winning

Maurizio Mancini<sup>1</sup>, Laurent Ach<sup>2</sup>, Emeline Bantegnie<sup>2</sup>, Tobias Baur<sup>3</sup>,  
Nadia Berthouze<sup>4</sup>, Debajyoti Datta<sup>5</sup>, Yu Ding<sup>5</sup>, Stéphane Dupont<sup>6</sup>,  
Harry J. Griffin<sup>4</sup>, Florian Lingenfeller<sup>3</sup>, Radoslaw Niewiadomski<sup>1</sup>,  
Catherine Pelachaud<sup>5</sup>, Olivier Pietquin<sup>7</sup>, Bilal Piot<sup>7</sup>, Jérôme Urbain<sup>6</sup>,  
Gualtiero Volpe<sup>1</sup>, and Johannes Wagner<sup>3</sup>

<sup>1</sup> InfoMus - DIBRIS, Università Degli Studi di Genova, Viale Francesco Causa, 13,  
16145 Genova, Italy

{maurizio.mancini,gualtiero.volpe}@unige.it,  
radoslaw.niewiadomski@dibris.unige.it

<sup>2</sup> LA CANTOCHE PRODUCTION, Hauteville, 68, 75010 Paris, France

{lach,ebantegnie}@cantoche.com

<sup>3</sup> Institut für Informatik, Universität Augsburg, Universitätsstr. 6a, 86159 Augsburg,  
Germany

{tobias.baur,florian.lingenfeller,  
johannes.wagner}@informatik.uni-augsburg.de

<sup>4</sup> UCLIC, University College London, Gower Street, London, WC1E 6BT,  
United Kingdom

{nadia.berthouze,harry.griffin}@ucl.ac.uk

<sup>5</sup> CNRS - LTCI UMR 5141 - Telecom ParisTech, 37-39 Rue Dareau, 75014 Paris,  
France

{debajyoti.datta,yu.ding,catherine.pelachaud}@telecom-paristech.fr

<sup>6</sup> TCTS Lab, Faculté Polytechnique, Université de Mons, Place du Parc 20, 7000  
Mons, Belgium

{stephane.dupont,jerome.urbain}@umons.ac.be

<sup>7</sup> SUPELEC / UMI 2958 GT-CNRS, Rue Edouard Belin, 2, 57340 Metz, France  
{olivier.pietquin,Bilal.Piot}@supelec.fr

**Abstract.** Developing virtual characters with naturalistic game playing capabilities is an increasingly researched topic in Human-Computer Interaction. Possible roles for such characters include virtual teachers, personal care assistants, and companions for children. Laughter is an under-investigated emotional expression both in Human-Human and Human-Computer Interaction. The EU Project ILHAIRE, aims to study this phenomena and endow machines with laughter detection and synthesis capabilities. The *Laugh when you're winning* project, developed during the eNTERFACE 2013 Workshop in Lisbon, Portugal, aimed to set up and test a game scenario involving two human participants and one such virtual character. The game chosen, the yes/no game, induces natural verbal and non-verbal interaction between participants, including frequent hilarious events, e.g., one of the participants saying “yes” or “no” and so losing the game. The setup includes software platforms, developed by the ILHAIRE partners, allowing automatic analysis and fusion of human participants’ multimodal data (voice, facial expression, body movements, respiration) in real-time to detect laughter. Further, virtual

characters endowed with multimodal skills were synthesised in order to interact with the participants by producing laughter in a natural way.

**Keywords:** HCI, laughter, virtual characters, game, detection, fusion, multimodal.

## 1 Introduction

Computer-based characters play an ever-increasing role in Human-Computer Interaction, not only for entertainment but also for education, as assistants and potentially in healthcare. Such emotionally complex interactions demand avatars that can detect and synthesise emotional displays. Laughter is a ubiquitous and complex but under-investigated emotional expression. The *Laugh when you're winning* eNTERFACE 2013 Workshop project builds on the work of the EU Project ILHAIRE<sup>1</sup> and on the previous eNTERFACE projects *AVLaughterCycle* [52] and *Laugh Machine* [33].

The project consists of an avatar actively participating in social games, in particular the *yes/no* game scenario. The avatar capabilities developed for game playing will have many applications beyond simple entertainment. The complex human-avatar interaction of a game demands considerable behavioural naturalness for the avatar to be a credible, trustworthy character. The avatar responds to user laughter in a highly customised way by producing laughter of its own.

Laughter detection and analysis among the speech, noise and body movements that occur in social games is achieved through multimodal laughter detection and analysis of audio, video, body movements and respiration. Laughter decisions integrate output from a module that drives mimicry behaviour, in response to the detected parameters of users' laughter, e.g., intensity.

The close interaction of a game scenario, proposed here, demands precise laughter detection and analysis and highly natural synthesised laughter. The social effect of avatar laughter also depends on contextual factors such as the task, verbal and nonverbal behaviours beside laughter and the user's cultural background [2,3]. In addition social context and emotional valence have been shown to influence mimicry [5]. Therefore, in a game scenario with both positive and negative emotions, laughter and mimicry must be well-implemented in order to enhance rather than inhibit interaction.

In the last part of the report we present an experiment, carried out during the eNTERFACE 2013 Workshop, which assesses users' perception of avatar behaviour in the direct interaction involved in the game scenario. The level of emotional response displayed by the avatar is varied: no response, responsive, responsive with mimicry. Measures of users' personality are analysed alongside short-term measures, e.g., user laughter, and long-term measures of engagement, e.g., mood, trust in the avatar. This spectrum of measures tests the applicability of an emotionally sensitive avatar and how its behaviour can be optimised to

---

<sup>1</sup> <http://www.ilhaire.eu>

appeal to the greatest number of users and avoid adverse perceptions such as a malicious, sarcastic or unnatural avatar.

## 2 Background

The concept of a game playing robot has long intrigued humans, with examples, albeit fake, such as the Mechanical Turk in the 18th century [38]. Games are complex social interactions and the possibility of victory or defeat can make them emotionally charged. The importance of emotional competence (the ability to detect and synthesise emotional displays) has therefore been recognised in more recent human-avatar/robot systems. Leite et al. [25] describe an empathic chess-playing robot that detected its opponent's emotional valence. More children reported that the robot recognised and shared their feelings when it displayed adaptive emotional expressions during games.

Laughter often occurs in games due to their social context and energetic, exhilarating nature. Recognising and generating laughter during games is therefore vital to an avatar being an engaging, emotionally convincing game companion. In addition, a trend for gamification - "the use of game attributes to drive game-like behavior in a non-gaming context may increase emotional expressions, such as laughter, in serious or mundane tasks" [35]. Thus an emotionally competent avatar developed for a game situation may well have value in situations such as education, exercise or rehabilitation.

## 3 State of the Art

### 3.1 Laughter Installations

Previous laughter detection and response systems have generally used a limited human-avatar interaction. Fukushima et al. [15] built a system that enhanced users' laughter activity during video watching. It comprised small dolls that shook and played prerecorded laughter sounds in response to users' laughter.

AVLaughterCycle aimed to create an engaging laughter-driven interaction loop between a human and the agent [52]. The system detected and responded to human laughs in real time by recording the user's laugh and choosing an acoustically similar laugh from an audiovisual laughter database.

The Laugh Machine project endowed a virtual agent with the ability to laugh with a user as a fellow audience member watching a funny video [53,33]. The agent was capable of detecting the participant's laughs and laughing in response to both the detected behaviour or to pre-annotated humorous content of the stimulus movie. The system was evaluated by 21 participants taking part in one of three conditions: interactive laughter (agent reacting to both the participant's laughs and the humorous movie), fixed laughter (agent laughing at predefined punchlines of the movie) or fixed speech (agent expressing verbal appreciation at predefined punchlines of the movie). The results showed that the interactive agent led to increased amusement and felt contagion.

### 3.2 Laughter Detection

Laughter has long been recognised as a whole-body phenomenon which produces distinctive body movements. Historical descriptions of these movements include bending of the trunk, movement of the head and clutching or slapping of the abdomen or legs [40]. The distinctive patterns of respiration that give rise to the equally distinctive vocalisations of laughter also generate movements of the trunk. An initial rapid exhalation dramatically collapses the thorax and abdomen and may be followed by a series of smaller periodic movements at lower volume. Fukushima et al. used EMG signals reflecting diaphragmatic activity involved in this process to detect laughter [15]. These fundamental laughter actions also drive periodic motion elsewhere in the body. Motion descriptors based on energy estimates, correlation of shoulder movements and periodicity to characterise laughter have been investigated [29]. Using a combination of these measures a Body Laughter Index (BLI) was calculated. The BLIs of 8 laughter clips were compared with 8 observers' ratings of the energy of the shoulder movement. A correlation, albeit weak, between the observers' ratings and BLIs was found. This model is used in the current project (see Section 6.2).

A body of work on recognition of emotion from body movements has accumulated in recent years [20,21,9,4,30]. Some of this work has concentrated on differences in movements while walking. Analysing the body movements of laughter presents a contrasting challenge in that, unlike walking, its emotional content cannot be modelled as variations in a repeated, cyclical pattern. Furthermore, body movements related to laughter are very idiosyncratic. Perhaps because of this, relatively little detection of laughter from body movements (as opposed to facial expressions) has been undertaken. Scherer et al. [43] applied various methods for multimodal recognition using audio and upper body movements (including head). Multimodal approaches actually yielded less accurate results than combining two types of features from the audio stream alone. In light of these results there is obviously considerable room for improvement in the contribution of body-movement analysis to laughter detection.

Discrimination between laughter and other events (e.g., speech, silence) has for a long time focused only on the audio modality. Classification typically relies on Gaussian Mixture Models (GMMs) [47], Support Vector Machines (SVMs) [47,19], Multi-Layer Perceptrons (MLPs) [22] or Hidden-Markov Models (HMMs) [8], trained with traditional spectral and prosodic features (MFCCs, PLP, pitch, energy, etc.). Error rates vary between 2 and 15% depending on the data used and classification schemes. Starting from 2008, Petridis and Pantic enriched the so far mainly audio-based work in laughter detection by consulting audio-visual cues for decision level fusion approaches [36]. They combined spectral and prosodic features from the audio modality with head movement and facial expressions from the video channel. They reported a classification accuracy of 74.7% in distinguishing three classes: unvoiced laughter, voiced laughter and speech [37]. Apart from this work, there exists to our knowledge no automatic method for characterizing laughter properties (e.g., emotional type, arousal, voiced or not). It must also be noted that few studies have investigated

the segmentation of continuous streams (as opposed to classifying pre-segmented episodes of laughter or speech) and that performance in segmentation is lower than classification performance [37].

### 3.3 Laughter Acoustic Synthesis

Until recently, work on the acoustic synthesis of laughter has been sparse and of limited success with low perceived naturalness. We can for example cite the interesting approach taken by Sundaram and Narayanan [44], who modeled the rhythmic energy envelope of the laughter acoustic energy with a mass-spring model. A second approach was the comparison of articulatory synthesis and diphone concatenation done by Lasarczyk and Trouvain [24]. In both cases the synthesized laughs were perceived as significantly less natural than human laughs. Recently, HMM-based synthesis, which had been efficiently applied to speech synthesis [46], has advanced the state-of-the-art [49].

### 3.4 Laughter Synthesis with Agents

Few visual laughter synthesis models have been proposed so far. The major one is by Di Lorenzo et al. [13] who proposed an anatomically inspired model of upper body animation during laughter. It allows for automatic animation generation of the upper-body from a preregistered sound of laughter. Unfortunately it does not synthesize head and facial movement during laugh. Conversely, a model proposed by Cosker and Edge [11] is limited to only facial animation. They built a data-driven model for non-speech related articulations such as laughs, cries etc. It uses HMM trained from motion capture data and audio segments. For this purpose, the number of facial parameters acquired with an optical motion capture system Qualisys was reduced using PCA, while MFCC was used for the audio input. More recently Niewiadomski and Pelachaud [32] have proposed a model able to modulate the perceived intensity of laughter facial expressions. For this purpose, they first analysed the motion capture data of 250 laughter episodes annotated with 5-point intensity scale and then extracted a set of facial features that are correlated with the perceived laughter intensity. By controlling these features the model modulates the intensity of displayed laughter episodes.

## 4 Innovation: Multimodality

As already explained, the *Laugh When You're Winning* project builds upon the Laugh Machine project that was carried out during eNTERFACE'12. The major innovations with regards to Laugh Machine or other installations are the following (these innovations will be further detailed in Sections 6 and 7):

- The laughter detection module has been extended to take multimodal decisions: estimations of the likelihoods of Smile, Speaking and Laughter likelihoods result from analyses of audio, facial and body movements, while in



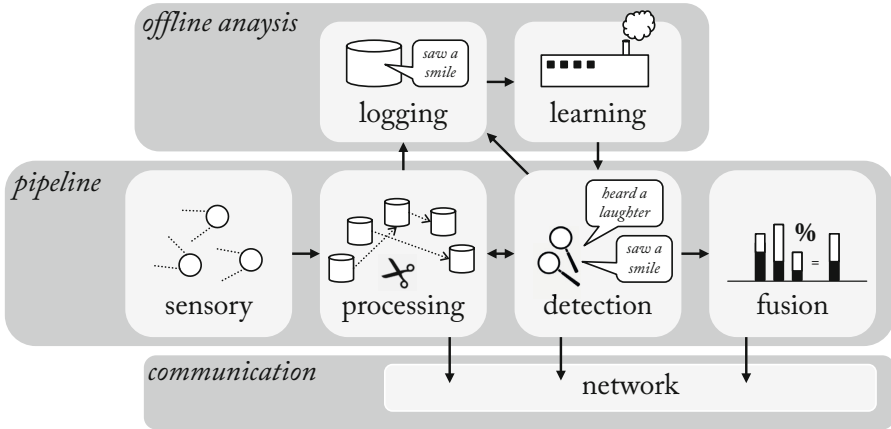
Laugh Machine there was simply a laughter/no-laughter detection based on audio only; in addition, the intensity estimation module has been improved (a neural network was trained under Weka).

- Several modules exchange information in real time for detection and analysis of laughter: the master process is Social Signal Interpretation (SSI) but some computations are outsourced to Eyesweb and Weka.
- The new game scenario, which can foster different types of emotions and involves 2 users simultaneously taken into account by the system; an ad hoc game engine has been developed to manage this specific scenario.
- The integration of laughter mimicry, through modules that analyse some laughter properties of (one of) the participants (e.g., shoulder movements periodicity) to influence the laughs displayed by the agent (shoulder periodicity and rhythm of the acoustic signal are driven by the measured properties).

## 5 Social Signal Interpretation (SSI)

The recognition component has to be equipped with certain sensors to capture multimodal signals. First, the raw sensor data is collected, synchronized and buffered for further processing. Then the individual streams are filtered, e.g. to remove noise, and transformed into a compact representation by extracting a set of feature values from the time- and frequency space. In this way the parameterized signal can be classified by either comparing it to some threshold or applying a more sophisticated classification scheme. The latter usually requires a training phase where the classifier is tuned using pre-annotated sample data. The collection of training data is thus another task of the recognition component. Often, activity detection is required in the first place in order to identify interesting segments, which are subject to a deeper analysis. Finally, a meaningful interpretation of the detected events is only possible in the context of past events and events from other modalities. For instance, detecting several laughter events within a short time frame increases the probability that the user is in fact laughing. On the other hand, if we detect that the user is talking right now we would decrease the confidence for a detected smile. The different tasks the recognition component undertakes are visualized in Figure 1.

The Social Signal Interpretation (SSI) software [54] developed at Augsburg University suits all mentioned tasks and was therefore used as a general framework to implement the recognition component. SSI provides wrappers for a large range of commercial sensors, such as web/dv cameras and multi-channel ASIO audio devices, or the Microsoft Kinect, but other sensors can be easily plugged to the system thanks to a patch-based architecture. It also contains processing modules to filter and/or extract features from the recording signals. In addition, it includes several classifiers (K-nearest Neighbor, Support Vector Machines, Hidden Markov Models, etc.) and fusion capabilities to take unified decisions from several channels.



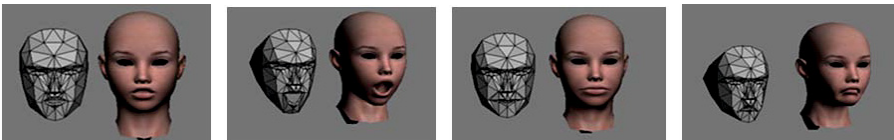
**Fig. 1.** Scheme of the laughter recognition component implemented with the Social Signal Interpretation (SSI) framework. Its central part consists of a recognition pipeline that processes the raw sensory input in real-time. If an interesting event is detected it is classified and fused with previous events and those of other modalities. The final decision can be shared through the network with external components.

In this project, SSI was used to synchronize the data acquisition from all the involved sensors and computers, estimate users' states (laughing, speaking or smiling) from audio (see Laugh Machine project [33]) and face (see Section 6.1, as well as fusing the estimations of users' states coming from the different modalities: audio, face analysis and body analysis ((outsourced to Eyesweb, see Section 6.2).

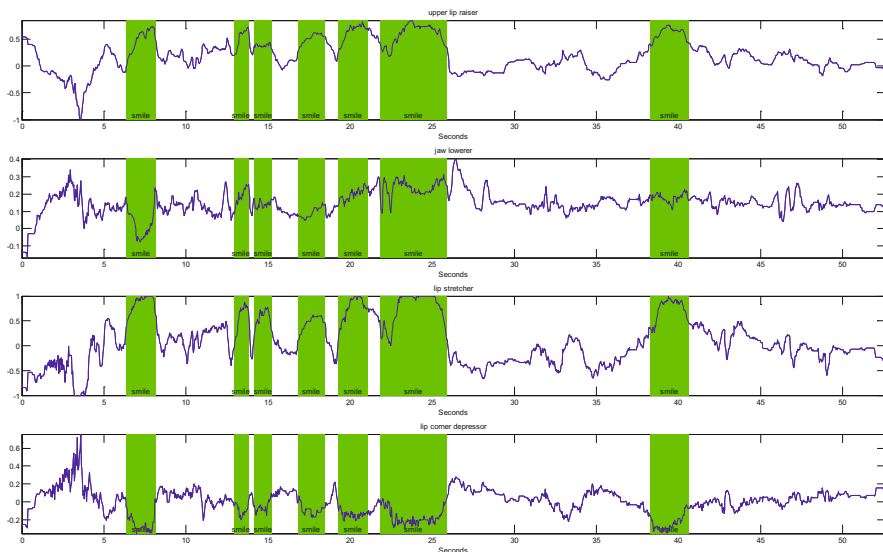
## 6 Multimodal Laughter Detection

### 6.1 Face Analysis

Face tracking provided by the Kinect SDK gives values for 6 action units (AUs) that are used to derive the probability that the user is smiling (in particular position of the upper lip and lip corners). In our tests we selected 4 of them as promising candidates for smile detection, namely *upper lip raiser*, *jaw lowerer*, *lip*



**Fig. 2.** Promising action units for smile detection provided by Kinect face tracking, namely *upper lip raiser*, *jaw lowerer*, *lip stretcher*, *lip corner depressor*



**Fig. 3.** Correlation between the measured action units and periods of laughter (green)

*stretcher*, *lip corner depressor* (see Figure 2). In order to evaluate these features test recordings were observed and analysed in Matlab.

Plots of the features over time are visualized in Figure 3. Laughter periods are highlighted in green. We can see that especially the values received for *upper lip raiser* (1st graph) and *lip stretcher* (3rd graph) are significantly higher during laughter periods than in-between laughter periods; *lip corner depressor*, on the other hand, has a negative correlation, i. e. values decrease during periods of laughter.

In order to combine the action units to a single value we found the following formula to give reasonable good results:

$$p_{smile} = upper\ lip\ raiser \times lip\ stretcher \times (1 - lip\ corner\ depressor) \quad (1)$$

In order to filter out natural trembling we additionally define a threshold  $T_{smile}$ . Only if above the threshold,  $p_{smile}$  will be included in the fusion process (see Section 6.3). In our test  $T_{smile} = 0.5$  gave good results.

As a second source of information Fraunhofer's tool SHORE [42] is used to derive a happy score from the currently detected face. Tests have shown that the happiness score highly correlates with user smiles. Including both decisions improves overall robustness.

## 6.2 Body Analysis

Real-time processing of body (i.e., trunk, shoulders) features is performed by EyesWeb XMI [27]. Compressed (JPEG) Kinect depth image streams captured

by SSI are sent on-the-fly via UDP packets to a separate machine on which EyesWeb XMI programs (called *patches*) detect shoulder movements and other body-movement measures, e.g., Contraction Index. Additionally, color-based tracking of markers (green polystyrene balls) placed on the user's shoulders is performed by EyesWeb XMI and the resulting recognition data is sent back to SSI to be integrated in the following overall laughter detection and fusion process (Section 6.3).

The body detection algorithms we present in this report are an improvement and extension of the techniques developed for the Laugh Machine (eNTERFACE'12) [33]. In particular, the previously described Body Laughter Index (BLI) is computed as a weighted sum of user's shoulders correlation and energy:

$$BLI = \alpha \bar{\rho} + \beta \bar{E} \quad (2)$$

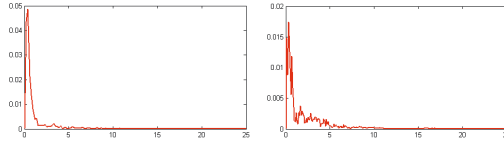
where the correlation  $\rho$  is computed as the Pearson correlation coefficient between the vertical position of the user's left shoulder and the vertical position of the user's right shoulder; and kinetic energy  $E$  is computed from the speed of user's shoulders and their mass relative to body mass.

We also validate the BLI by the user's shoulder movement frequency: if frequency is included in an acceptable interval  $[2, 8]Hz$  then the BLI is valid. The interval is motivated by psychological studies on laughter by Ruch and Ekman [40].

In this report we introduce a new information for the body (i.e., trunk, shoulders) modality: laughter intensity. When a laughter event is detected by using the BLI, the FFT of the Y component of shoulders and trunk is computed along the entire event length (events lasted from 1 second to 9 seconds). The two most prominent peaks of the FFT, *max1* (the absolute maximum) and *max2* (the second most prominent peak) are then extracted. These are used to compute the following index:

$$R = \frac{max1 - max2}{max1} \quad (3)$$

Basically, the index will tend to 1 if just one prominent component is present; it will tend to 0 if two or more prominent components are present. Thus, periodic movements, i.e., those exhibiting one prominent component, will be characterized by an index near 1, while the index for non-periodic movements will be near 0. Figure 4 shows two examples of such computation: on the left, one peak around 1/3 Hz is shown, probably related to torso rocking during laughter, and the index tends to be close to 1, indicating a highly periodic movement; on the right, many peaks between 1/3 Hz and 1.5 Hz are shown, and the index is close to 0, indicating a mildly periodic movement.



**Fig. 4.** FFT computed on the shoulder Y coordinate. On the left a prominent component is present and the index tends to 1. On the right many prominent components are present and the index tends to 0.

A test carried out in the past months on 25 laughter events annotated for intensity by psychologists [28], showed that  $R$  can successfully approximate laughter intensity. Significant correlations between  $R$  and the manually annotated intensity values were found for both shoulders ( $r = 0.42, p = 0.036$ ) and trunk ( $r = 0.55, p = 0.004$ ).

**Table 1.** Correlation between body indexes and annotated laughter intensity

Index	Correlation	p-Value
$R_s$	0,4216	0,0358
$R_t$	0,5549	0,0040
$R_d$	0,1860	0,3732

Table 1 reports correlation and p-values for shoulder/trunk indexes and annotated laughter intensity.  $R_s$  is the index computed only on shoulder movement;  $R_t$  is the same index computed only on trunk movement;  $R_d$  is the index computed on the difference between shoulder and trunk movement (that is, shoulder movement relative to trunk position).

### 6.3 Detection and Fusion

During fusion a newly developed event based vector fusion enhances the decision from the audio detector (see [33]) with information from the mentioned sources. Since the strength of the vectors decays over time, their influence on the fusion process decreases, while they still contribute to keep recognition stable. The final outcome consists of three values expressing probability for talking, laughing and smiling.

The method is inspired by work by Gilroy et al. [16] and adapted to the general problem of fusing events in a single or multi-dimensional space. A main difference from their proposed method is that a modality is not represented by a single vector, but new vectors are generated with every new event. In preliminary experiments this led to much more stable behaviour of the fusion vector, since the influence of false detections is lowered considerably. Only if several successive events are pointing in the same direction is the fusion vector dragged into this direction.

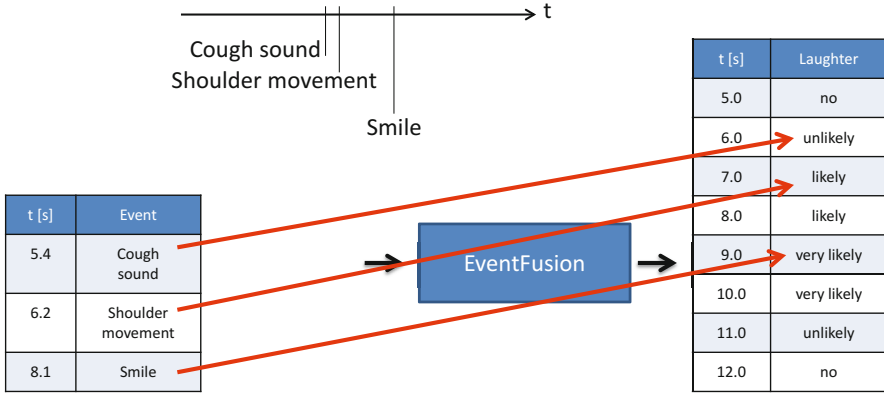


Fig. 5. Example of event fusion

The algorithm is illustrated in Figure 5. In the example three successive events are measured: a cough sound, a shoulder movement shortly after that and, after a small delay, a smile. Each event changes the probability that the user is laughing. When the first event, the cough sounds, arrives it is still unlikely, since, although it is a vocal production, coughing differs from laughter. However, a shoulder movement is detected shortly after, laughter becomes more likely and the laughter probability is increased. And when finally a smile is detected the laughter probability becomes even more likely. Due to the decay function that is applied to the event vectors the probability afterwards decreases over time.

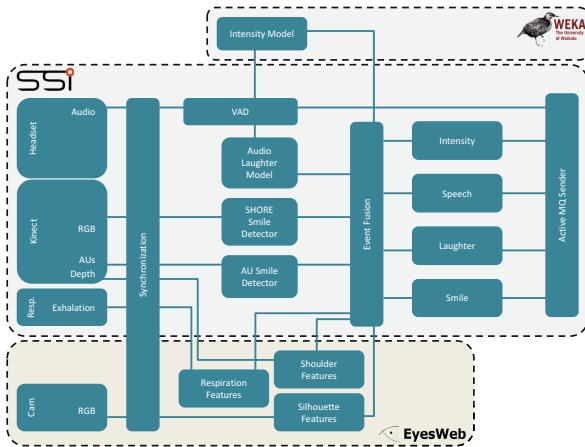


Fig. 6. The final detection system

Thanks to the fusion, performance in terms of reliability and robustness has clearly been improved compared to the previous system. A schema of the final detection system is shown in Figure 6.

## 7 Multimodal Laughter Synthesis

### 7.1 Dialogue Manager

The original objective of the project was to train a dialogue manager from human data; however, this component could not be built within the time constraints of the project. To allow for the interaction to take place, a rule-based dialogue manager with empirical thresholds was designed. It follows simple rules to decide when and how (duration, intensity) the agent should laugh, given the state of the game (game in progress, game finished) and the detected states of the two participants (speaking, laughing, smiling or none of these). The implemented rules are presented in Table 2. Empirical thresholding on the speaking, laughing and smiling likelihoods was used to determine the state of each participant. The implemented rules are symmetric with respect to the two participants (no difference is made between speaker and observer, the same rules apply if the participants are switched).

**Table 2.** Rules for determining when and how the agent should laugh. The implemented rules are symmetric (Participant1 and Participant2 can be reversed). If several rules are met (e.g. likelihoods for Laughter and Speech of Participant 1 both reach the defined thresholds, the highest rule in the table receives priority.

Participants states		Laughter decision		Participants states		Laughter decision	
P1	P2	Intensity	Duration	P1	P2	Intensity	Duration
Laugh	Laugh	High	High	Speak	Smile	Low	Low
Laugh	Speak	Low	Low	Speak	Silent	/	/
Laugh	Smile	Medium	Medium	Smile	Smile	Medium	Medium
Laugh	Silent	Medium	Medium	Smile	Silent	Low	Low
Speak	Speak	/	/	Silent	Silent	/	/

The dialog manager also considers the game context, which is obtained thanks to mouse clicks send by SSI. A click on Mouse 1 signals the start of a yes/no game. A click on Mouse 2 tells that the speaker has lost the game (by saying “yes” or “no”). Thanks to these clicks, the dialog manager can determine at every moment the state of the game, which can take 4 different values: game not started yet, game on, game lost, game won<sup>2</sup>. This information on the game state is further transmitted to the laughter planner.

### 7.2 Laughter Planner

Laughter Planner controls the behavior of the agent as well as the flow of the game. It decides both verbal and nonverbal messages taking into account the

<sup>2</sup> In our case, the game is won if the speaker manages to avoid saying yes or no during 1 minute, so the dialog manager puts the game status to game won one minute after the game started (click on Mouse 1), if there was no click on Mouse 2 (game lost) in the meantime.

**Table 3.** Laughter Planner Inputs (DM = Dialog Manager; MM = Mimicry Module)

Name	Description	Values	Sender
LAUGH_DUR	Duration of the laugh to be displayed by the agent	R+	DM
LAUGH_INT	Intensity of the laugh to be displayed by the agent	[0, 1]	DM
MIMICKED_AMP	relative amplitude of human laughter	[-1, 1]	MM
MIMICKED_VEL	relative velocity of human laughter	[-1, 1]	MM
SPEECH_P_SPR	probability that the speaker is currently speaking	[0, 1]	SSI
SPEECH_P_OBS	probability that the observer is currently speaking	[0, 1]	SSI

verbal and nonverbal behavior of the human participants of the game, who are denoted the speaker (SPR; the person that is challenged in the game) and the observer (OBS; i.e. the second human player that also poses the questions), and the rules of the game. Laughter Planner receives continuously the inputs presented in Table 3.

The main task of Laughter Planner is to control the agent behavior. The details of the algorithm are presented in Figure 7. Laughter Planner generates both the questions to be posed by the agent to the human player as well as laughter responses.

The game context is also taken into account: the agent is only allowed to ask questions when the game is on; when the game is won, the agent informs the participants (e.g., “Congratulations, the minute is over, you won!”); when the game is lost, the agent laughs in the laughter conditions, or says something in the no-laughter condition e.g., “Oh no, you just lost”).

The questions are selected from the pre-scripted list of questions. This list contains the questions that were often used by humans when playing the game (e.g. MMLI corpus [31]). Some of the questions are independent of others while others are asked only as part of a sequence e.g. “What’s your name?” ... “Is that your full name” ... “Are you sure?”. The questions in sequence should be displayed in the predefined order, while the order of other questions is not important. Additionally, Laughter Planner takes care not to repeat the same question twice in one game. Each question is coded in BML that is sent at the appropriate moment to any of two available Realizers (Greta, Living Actor). The Planner poses a new question when neither of the human participants speak for at least 5 seconds. If the observer starts to speak, he probably poses a new question or a new sequence of questions. In that case, Laughter planner abandons its sequence of questions and starts a new one in the next turn.

Also the set of laughter is predefined. The laughter episodes (audio and facial expressions) are pre-synthesized off-line (see Sections 7.4 and 7.5 for details) from the available data (AVLC corpus). Only the shoulder movements are generated in real time. For each episode of AVLC corpus, five different versions were created, each of them with different laugh burst duration and consequently also different durations. Thus each original sample can be played “quicker” or “slower” and also corresponding lip movement animation and shoulder movement animation are accordingly modified. All the pre-synthesized laughs are



divided into 3 clusters according to their duration and intensity. Additionally each cluster is divided into 5 subclusters according to the mean laugh burst velocity. While the choice of the episode is controlled with 2 input parameters sent by Dialog Manager (see Table 3), the 2 parameters sent by Mimicry Module are used to choose the correct velocity variation of the episode. In more details, the values of LAUGH\_DUR and LAUGH\_INT are used to choose a cluster of laugh episodes. Next, mimicry parameters are used to choose a subcluster of this cluster of episodes, i.e. a set of laughs of laugh bursts corresponding to the value sent by mimicry module. Finally, the laugh episode is chosen randomly from the subcluster and BML messages containing the name of episode as well as BML tags describing the animation over different modalities are sent to the Realizer.

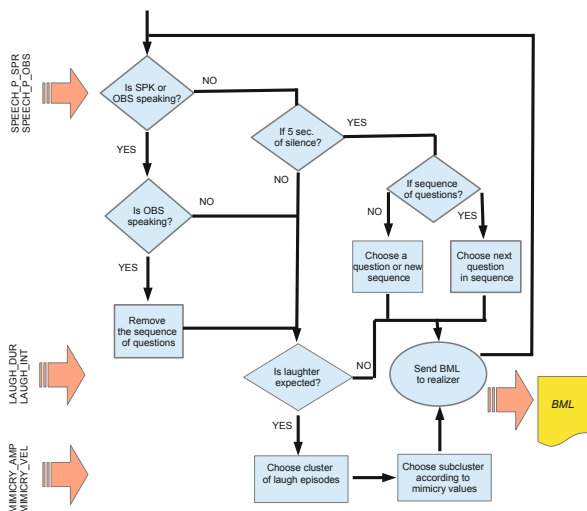


Fig. 7. Laughter Planner

### 7.3 Mimicry

The mimicry module has the task of deciding how the agent's expressive behaviour should mimic the user's one. The Greta agent has the capability to modulate its quality of movement (e.g., amplitude, speed, fluidity, energy, etc) depending on a set of *expressivity parameters*.

As illustrated in Figure 8, the mimicry module receives a vector of the user's body movement features  $X$  (see Section 6.2) as well as laughter probability ( $FLP$ ) and intensity ( $FLI$ ) resulting from the modality fusion process (see Section 6.3).

The mimicry module starts to work in *non-laugh* state. When  $FLP$  passes a fixed threshold  $T_1$ , the mimicry module enters the *laugh* state and starts to accumulate body features in  $X_E$ . In order to avoid continuous fast switching

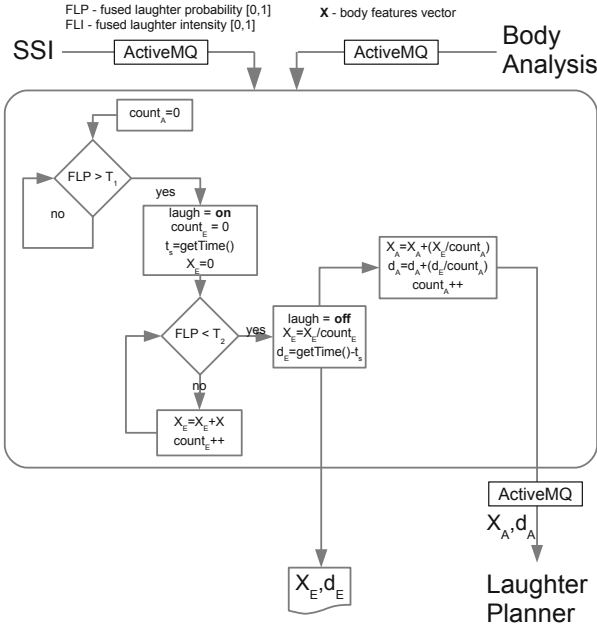


Fig. 8. Mimicry module

between laugh and non-laugh state,  $FLP$  is then compared against a second, lower, threshold  $T_2$ . When  $FLP$  goes under this threshold the mimicry module goes back to the non-laugh state. This means that the laughter event ends and a few operations are performed:

- the vector of the laughter event mean body features is computed as the ratio between  $X_E$  and the duration, in frames, of the event  $count_E$ ;
- the duration of the event, in seconds,  $d_E$  is computed;
- the *overall* mean body features vector  $X_A$  is computed as the incremental mean of  $X_E$ ;
- the *overall* mean event duration  $d_A$  is computed as the incremental mean of  $d_E$ ;
- the mean body features vector  $X_E$  and the event duration  $d_E$  are stored into a file for later offline use;

Finally, the overall mean body features vector  $X_A$  and event duration  $d_A$  are sent to the Laughter Planner (see Section 7.2) where they will contribute to modulate the agent’s laughter duration and body expressive features.

#### 7.4 Acoustic Laughter Synthesis

Acoustic laughter synthesis technology is the same as presented in [50]. It relies on Hidden Markov Models (HMMs) trained under HTS [34] on 54 laughs uttered by one female participant of the AVLaughterCycle recordings [52]. After

building the models, the same 54 laughs have been synthesized using as only input to the phonetic transcriptions of the laughs. The best 29 examples have been selected for the current experiments (the other 25 examples had disturbing or badly rendered phones due to limited number of the corresponding phones in the training data).

To increase the number of available laughs and the reactivity of the system, phonetic transcriptions of laughter episodes composed of several bouts (i.e., exhalation parts separated by inhalations) have been split into bouts by cutting the original transcription at the boundaries between inhalation and exhalation phases. This indeed increases the number of laughter examples (for example one episode composed of three bouts will produce three laughter segments instead of one). This method also increases reactivity of the system - which is limited by the impossibility of interrupting currently playing laughs - as shorter laughter segments are available: instead of the original episode of, for example, 15s, the repository of available laughter now includes three bouts of, for example, 6, 4 and 5s, which would “freeze” the application for a shorter time than the initial 15s.

To enable mimicry of the rhythm in the application, several versions of the laughs have been created: laughs are made rhythmically faster or slower by multiplying the durations of all the phones in the laughter phonetic transcription by a constant factor  $F$ . The laughs corresponding to the modified phonetic transcriptions are then synthesized through HTS, with the duration imposed to respect the duration of the phonetic transcription (in other words, the duration models of HTS are not used). Laughs have been created with this process for the following values of  $F$ : 0.6, 0.7, 0.8, 0.9, 1 (original phonetic transcription), 1.1, 1.2, 1.3 and 1.4.

Finally, the acoustically synthesized laughs are placed in the repository of available laughs, which contains for each laugh: a) the global intensity of the laugh, derived from the continuous intensity curve computed as explained in [48]; b) the duration of the laugh; c) the audio file (.wav); d) the phonetic transcription of the laughs, including the intensity value of each phone; e) the rhythm of the laugh, computed as the average duration of “fricative-vowel” or “silence-vowel” exhalation syllables of the laugh.

The first two pieces of information are used for selecting the laugh to play (using the clustering process presented in section 7.2). The next two (audio and transcription files) are needed by the agent to play the selected laugh. Finally, the rhythm of the laugh is used to refine the selection when mimicry is active (only laughs within a target rhythm interval are eligible at each moment).

## 7.5 Greta

### Facial Animation

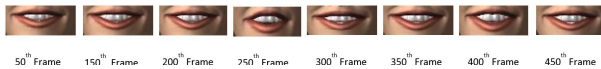
As for the audio signal (Section 7.4), our work is based on the AVLIC data set [52]. 24 subjects (9 women and 15 men) were recorded while watching humorous videos. This corpus includes 995 laughter examples: video, audio and facial

motion capture data. Laughs were phonetically annotated [51]. Automatic landmark localization algorithm was applied to all the laughter example videos for extracting the trajectories of Facial Animation Parameters (FAPs) (see [39]). In our model we use 22 lip FAPs as lip motion features, 3 head rotation FAPs as head features and 8 eyebrow FAPs as eyebrow features. Therefore, we have the lip, head and eyebrow motions and phonetic information of all the laughter examples included in AVLC.

Lip movements play an important role in human voice production. They are highly synchronized with spoken text, e.g., phoneme. Humans can easily perceive whether spoken text and visual lip motion are synchronized. Therefore, virtual agents should be capable of automatically generating believable lip motion during voice production. Phonetic sequences have been used to synthesize lip movements during speech in previous papers [6,10,7,23,14,12,26], most of which use the mapping between lip form (visual viseme) and spoken phoneme. To our knowledge, no effort has focused on natural synthesis of laughter lip motions.

One of our aims is to build a module that is able to automatically produce lip motion from phonetic transcriptions (i.e., a sequence of laughter phones, as used for acoustic synthesis). This work is based on the hypothesis that there exists a close relationship between laughter phone and lip shape. This relationship is learned by a statistical framework in our work. Then the learnt statistical framework is used to synthesize the lip motion from pseudo-phonemes and duration sequences.

We used a Gaussian Mixture Model (GMM) to learn the relationship between phones and lip motion based on the data set (AVLC). The trained GMM is capable of synthesizing lip motion from phonetic sequences. One Gaussian distribution function was learnt to model the lip movements for each of the 14 phonetic clusters used for laughter synthesis. Therefore, the trained GMM was comprised of 14 Gaussian distribution functions. For synthesis, one phonetic sequence including the duration of each phone is taken as the input, which is used to establish a sequence of Gaussian distribution functions. The determined sequence of Gaussian distribution functions [45] is used to synthesize directly the smoothed trajectories. Figure 9 shows an example of synthesized lip motion.



**Fig. 9.** Lip motion synthesized from a phonetic transcription

Head and eyebrow behaviours also play an important role in human communication. They are considered as auxiliary functions of speech for completing the human expressions. For example, they can convey emotional states and intentions. Humans are skilled in reading subtle emotion information from head and eyebrow behaviours. So, human-like head and eyebrow behaviour synthesis is necessary for a believable virtual agent. In consequence, we wanted to synthesize head and eyebrow motion in real time from the phonetic sequences. The

proposed approach is based on real human motions recorded in the database. All the motion data sequences in the database were segmented according to the annotated phonetic labels. The motion segments were categorized into 14 clusters corresponding to the 14 phonetic classes.

We developed a new algorithm for selecting an optimal motion segment sequence from the 14 motion segment clusters, according to the given phonetic sequence. In the proposed algorithm, one cost function is defined to evaluate the costs of all the motion segments belonging to the cluster corresponding to the given phonetic label. The cost value consists of two sub-cost functions. The first sub-cost called duration cost is the difference between the motion segment duration and the target duration; the second sub-cost called position cost is the position distance between the value of the first frame of the motion segment and the value of last frame of the previously selected motion segment. The motion segment with the smallest cost value is selected.

## Shoulder Movement

Previously the analysis of the motion capture data of the Multimodal Multi-person Corpus of Laughter in Interaction (MMLI) [31] has shown regularities in the shoulder movements during the laughter. In more detail, 2D coordinates of the shoulders' positions were processed using the Fast Fourier Transform (FFT). The results showed peaks in the frequency range  $[3, 6]Hz$ . Interestingly, from the analysis of acoustic parameters we know that similar frequencies were observed in audio laugh bursts [1,40]. Both these sources of information were used to generate shoulder movements that are synchronized with the synthesised audio (see Section 7.4).

The shoulder movements in the Greta agent are controlled by BML tags sent by Laughter Planner. The tag *shoulder* specifies the duration of the movement as well as its two additional characteristics: period and amplitude. These parameters are chosen by the Laughter Planner (see Section 7.2). In particular the period of the movement corresponds to the mean duration of the laugh burst in the laughter episode to be displayed. The amplitude of the shoulder movement corresponds to the amplitude of the movements detected within the Mimicry Module. If the detected movements are large then also the amplitude of the agent movements is higher, and conversely. Next, the shoulders' BML tags with all these parameters are turned into a set of frames. The vertical position of the shoulder joints is computed for each frame by using the following function:

$$X(t) = Amplitude * \cos(2 * PI * frequency * t - 75.75) \quad (4)$$

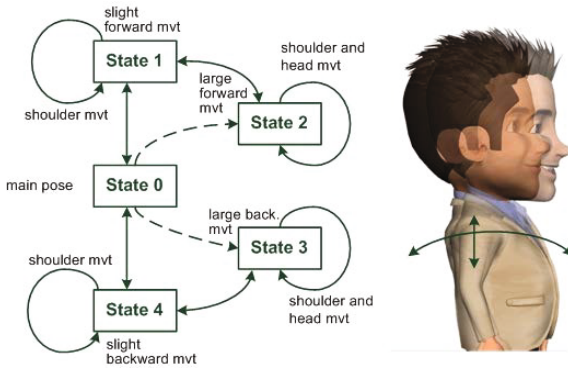
where the *amplitude* and *frequency* are parameters of the BML.

## 7.6 Living Actor<sup>TM</sup>

The Living Actor<sup>TM</sup> module includes a 3D real-time rendering component using Living Actor<sup>TM</sup> technology and a communication component that constitutes

the interface between the Living Actor<sup>TM</sup> avatar and the ActiveMQ messaging system. This version is based on sample animations created by 3D artists and combines “laughter” faces (facial expressions associated with visemes that are mouth movements corresponding to synthesized laughter sounds), “laughter” body animations corresponding to several types of movements (backward bending, forward bending, shoulder rotation) and “laughter” intensities. The main animation component is related to the trunk and arms that are combined with additional animations of head and shoulders.

The prepared trunk animations are later connected to form a graph so the avatar changes its key body position (State) using transition animations. The states in the graph (see Fig. 10) correspond to different types of laughing attitudes (bending forward, bending backward, shoulder rotations). Head and shoulder back-and-forth movements are not part of this graph; they are combined with graph transitions at run time. Some low amplitude animations of the arms are added to trunk animations so the avatar does not look too rigid.



**Fig. 10.** Sample laughter graph of animation

Living Actor<sup>TM</sup> software is originally based on graphs of animations that are combined with facial expressions and lips movements. Two main capabilities have been added to this mechanism:

- combine several animations of the body (torso, head, shoulder)
- use special facial expressions corresponding to laughter phones

The software is now able to receive data about phones and laughter intensity in real time. Depending on the received laughter intensity, a target state is chosen in the graph and transitions are followed along a path computed in real time. The input data, that include specific types of “laughter” movements, like bending forward or backward, are taken into account to choose the target states. Otherwise, one of the available types of movements is chosen by the avatar module, depending on intensity and random parameters.

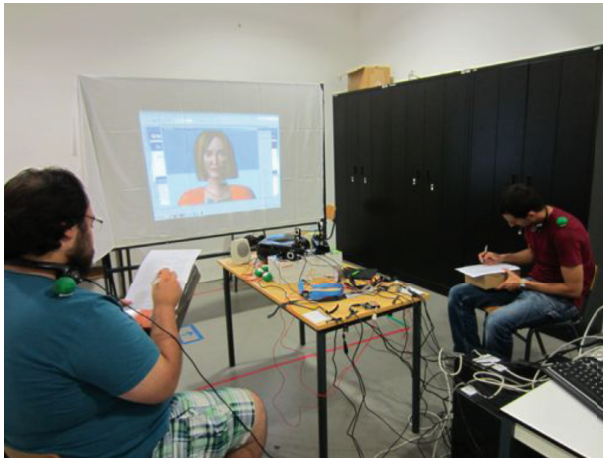
The animations triggered by the graph traversal are combined with head and shoulders back-and-forth movements that make the avatar “laughter” animations more realistic and avoid the perception of repetition when the same state is targeted several times in the graph. The data received from synthesized phonemes in real time are used to add facial morphing and lips movements.

When there is no instruction, the 3D real-time rendering component automatically triggers “Idle animations, so the avatar breathes, glances, or moves slightly and is never static.

## 8 Experiment

A preliminary experiment was run with the aim of evaluating the integrated architecture and the effect of the mimicry model on the participants. The avatar Greta was used for this experiment.

Eighteen participants (12 male, average age 26.8 (3.5) - 5 participants did not report their age) from the eNTERFACE workshop were recruited. They were asked to play the Yes/No game with the avatar in pairs. In the game, participants take turns in asking questions (observer) with the aim of inducing the other participant (speaker) to answer “yes” or “no”. Each turn lasted a maximum of 1 minute or until the participant answering the questions said “yes” or “no”. The avatar always played the role of supporting the observer by asking questions when a long silence occurred.



**Fig. 11.** Setting of the experiment. The participants are filling in an in-session questionnaire.

A within-subjects design was used: participants were asked to play the game in three different conditions: avatar talking but without exhibiting any laughter expression (No-Laughter condition), avatar exhibiting laughter expressions

(Laughter condition), avatar with laughter expression and long term mimicry capabilities (Mimicry condition). In all three conditions the avatar had laughter detection capabilities. In both the Laughter and the Mimicry conditions, the laughter responses were triggered by the detection of the laughter or smile in at least one of the participants (see Section 7.1). The order of the conditions was randomized. Each condition involved two turns of questioning, one for each participant.

The setting of the experiment is shown in Figure 11. The participants and the avatar sat around a table as shown in the figure. Each participant was monitored by a Microsoft Kinect and two webcams placed on the table. They were also asked to wear a custom made respiration sensor around their chest and a microphone around their neck.

Before the experiment, the participants had the game explained to them and were asked to sign a consent form. They were also asked to fill in a set of pre-experiment questionnaires:

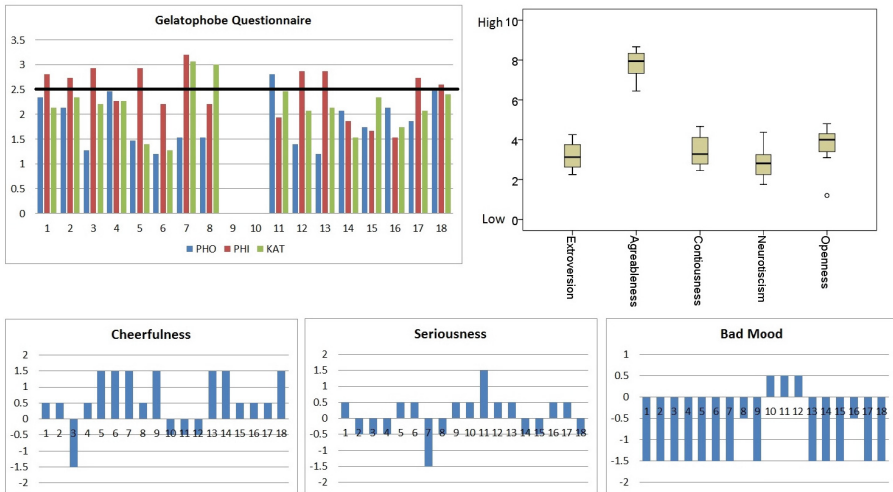
- “PhoPhiKat-45”: this provides scales to quantify levels of gelotophobia (the fear of being laughed at), gelotophilia (the joy of being laughed at), and katagelaticism (the joy of laughing at others) [41]. Questions are answered on a 4-point scale (1-4) and a person is deemed to have a slight expression of gelotophobia if their mean score is above 2.5 and pronounced gelotophobia if their mean score is greater than 3.
- A Ten Item Personality Inventory (TIPI): this measure is a 10-item questionnaire used to measure the five factor personality model commonly known as the “big five” personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [17].
- Current general mood: cheerfulness, seriousness and bad mood rated on a 4-point scale.
- Avatar general perception: this questionnaire measures the level of familiarity with, and the general likeability and perceived capability of avatars through a 8-item questionnaire.

After each condition, the participants were asked to fill in a questionnaire to rate their experience with the avatar (in-session questionnaire [18]). This questionnaire is a revised version of the LAIEF-R questionnaire developed for the evaluation experiment run at eNTERFACE’12. The new version includes questions about mimicry and body expression perception and is hereafter called LAIEF-Game.

At the end of the experiment, the participants were also asked to provide comments about the overall system [18]. Each experiment lasted about 1 hour.

A second round of four games was then played in one of the two remaining conditions (randomly assigned), followed by the same questionnaire answering. Then, a last round of four games was played in the remaining condition. Finally, the participants filled the interaction questionnaire as well as a general questionnaire.





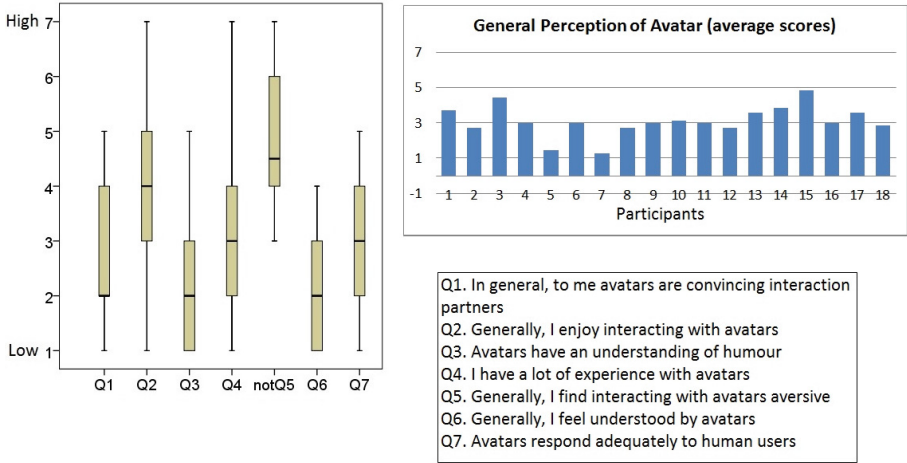
**Fig. 12.** (Top) Personality trait scores and (Bottom) current mood scores. X-axes indicate the participant number. Participants 9 and 10 did not fill in the personality traits questionnaires (Top). All participants filled in the current mood questionnaire (Bottom).

## Results

Figure 12 (top) shows the participants' personality traits in terms of gelatophobia and of extroversion, agreeableness, conscientiousness, neuroticism, and openness. Only 2 participants scored above the threshold for gelatophobia ( $PHO > 2.5$ ). The general mood (Figure 12 - bottom) was also measured as it could have an effect on the perception of the avatar during the experiment. The figure shows that the participants were overall in a good mood with only three participants scoring high in bad mood.

Figure 13 shows the level of familiarity with and the general likeability of avatars reported by our participants before starting the experiments. We can see from the boxplot for Q4 that our participants present a quite varied level of familiarity with avatars with most of them scoring in the lower part of the scale. The scores for the other questions are also quite low. Only Q2 ("Generally, I enjoy interacting with avatars") and Q5 ("Generally I find interacting with avatars aversive", score inverted for reporting) obtained quite high scores. This shows that, in general, our participants did not dislike interacting with avatars but they had a low confidence in the capabilities that avatars can exhibit when interacting with people.

In order to identify possible effect of laughter expression on the perception of the avatar, the questions from the in-session questionnaires were grouped into three factors: competence, likeability, naturalness. Naturalness was also separately explored with respect to: naturalness of the non-verbal expressions (excluding laughter-related questions) and of laughter expressions. The grouping of the questions was as follow:

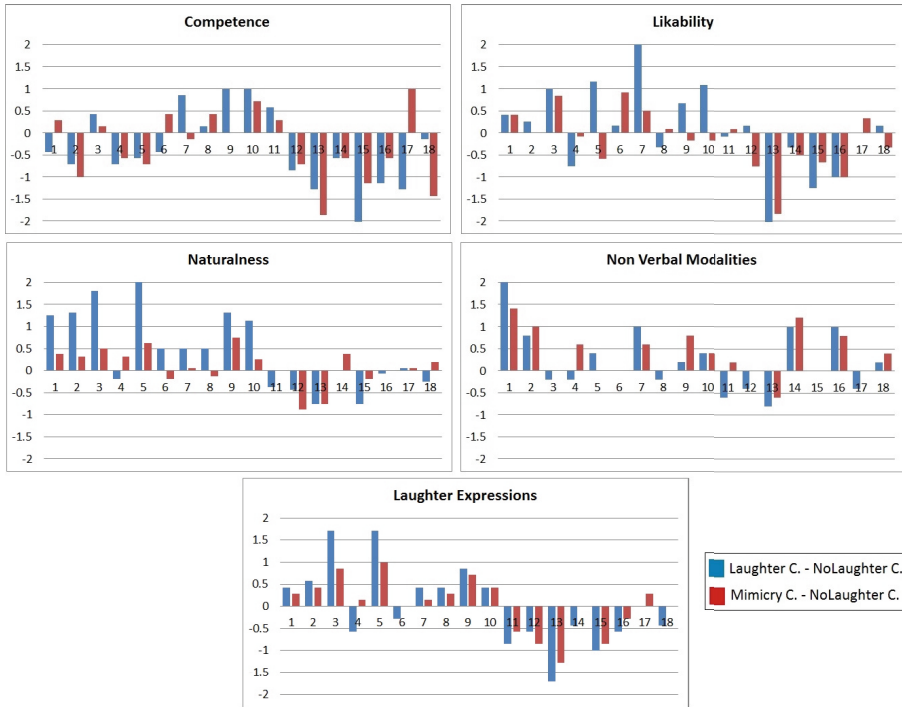


**Fig. 13.** General familiarity with and perception of likeability and competence of avatars. (Left) scores organized by question; (bottom-right) Q1-Q7 questions. “notQ5” indicates that the response has been inverted for reporting.; (top-right) average scores over the 7 questions for each participant.

- Competence: Q11, Q13, Q14, Q15, Q17, Q21, Q39
- Likeability: Q12, notQ16, Q18, notQ19, Q20, Q23, Q26, Q27, Q32, Q34, Q35, Q36
- Naturalness: Q22, Q25, Q31, Q37, Q38, Q40, Q41, Q42, Q47, NV, LN (excluding Q24, Q28)
- Non-verbal expressions (NV): Q29, Q30, Q40, Q41, Q42
- Laughter naturalness (LN): Q24, Q28, notQ43, Q44, Q45, Q46

Q24 and Q28 were excluded from the Naturalness factor since many participants did not answer these two questions for the no-laughter condition. These questions were however included in the laughter naturalness factor and a baseline value of 3.5 (middle of the scale) was used when the participant’s score was missing.

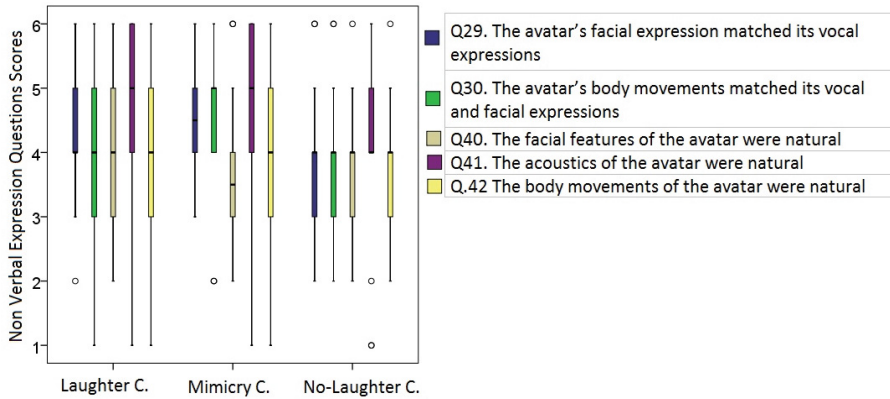
The list of questions can be seen in [18]. For each of these factors the scores were normalized. The differences between the laughter condition scores and the no laughter condition scores are shown in Figure 14. The data show high variability between participants’ scores. However, some trends can be identified. In particular, the avatar was perceived as a more competent game player in the control conditions than in any of the two conditions with laughter expressions. In the case of likeability, there is a clear split in the participants’ reaction to the avatar with many participants reporting greatly increased or decreased liking of the avatar in the laughter conditions compared to the control conditions. A more positive effect is observed in term of naturalness of the avatar.



**Fig. 14.** Comparison of in-session scores between conditions. X-axes indicate participant number; Y-axes indicate the differences between scores obtained in either the laughter condition or the mimicry conditions with respect to the control condition.

A repeated-measures test was run to investigate if there were any significant difference between the three conditions. Mauchly's test indicated that the assumption for sphericity was violated for naturalness ( $\chi^2(2) = 13.452, p < .01$ ), non-verbal expression naturalness ( $\chi^2(2) = 19.151, p < .01$ ) and laughter naturalness ( $\chi^2(2) = 9.653, p < .001$ ). Therefore a Greenhouse-Geisser correction was applied for these three factors. No significant effects were found for the perception of competence, likeability and laughter naturalness. However, significant effects were found for overall naturalness ( $F(1.178, 21.675) = 3.978, p = .05, \mu^2 = .190$ ) and of non-verbal expression ( $F(1.376, 23.4) = 4.278, p = .039, \mu^2 = .201$ ).

Post hoc comparisons for overall naturalness show that the laughter condition received higher scores than the other two conditions but these differences only approached significance (vs. no-laughter:  $p = .15$ ; vs. mimicry:  $p = 1.24$ ). Post hoc comparisons for non-verbal behaviour show a significant difference ( $p = 0.019$ ) between the no-laughter and mimicry conditions. Figure 15 shows the scores for each of the five questions forming the non-verbal expression factor. We can see that slightly higher scores were obtained for the laughter and mimicry condition with respect to the no-laughter condition. We can also observe higher scores for Q30 for the mimicry condition than for the laughter condition. It is possible that the greater amount of body behaviour (observed in the mimicry



**Fig. 15.** Boxplots of scores of the questions forming the non-verbal expression factor

condition) may have resulted in the avatar being perceived as more alive. It is also possible that the fact that, in the mimicry condition, the body behaviour was mimicking the body movement of the participants may have captured more their attention. However, only five participants reported feeling that the avatar was mimicking them and only 2 participants correctly indicated in which section the avatar was mimicking and which of the participants was mimicked. In addition, only one person reported that the avatar was mimicking their body movement.

The results of this first evaluation showed that laughter added some level of naturalness to the avatar; however, the evaluation also highlighted important technical and experimental design issues that will be addressed before running the full evaluation. In particular, because of the open audio production the avatar detected itself laughing and was unable to distinguish this from participant laughter, it then used this as a cue for generating ever-increasing laughter resulting at times in perceived random or hysterical laughter.

Some technical issues with the synthesis were also identified that need to be addressed to increase naturalness and facilitate communication (e.g., speech synthesis software). Comments from the participants were also very useful and highlighted different problems and solutions to address them. The scenario needs to be slightly redesigned to make sure that the position of the avatar in the triad is more central and participants do not exclude it from the game). Some Wizard of Oz techniques will be used to specifically evaluate individual modules of the laughter machine architecture (e.g., the mimicry module) to avoid the effect being masked by other technical issues (e.g., imperfect recognition of laughter, or lack of natural language understanding).

## 9 Conclusions

The *Laugh when you're winning* project was designed in the framework of the EU Project ILHAIRE, and its development took place during the eNTERFACE 2013

Workshop, where several partners joined to collaborate for the project setup. Further, the participation in the eNTERFACE Workshop allowed researchers to recruit participants for the testing phase. Tests showed that virtual characters laughter capabilities helped to improve the interaction with human participants. Further, some participants reported that they perceived whether the virtual character was mimicking their behavior.

Several critical points emerged from the project set up and testing and will be addressed in the future:

- the fused detection module is more robust than the one developed in eNTERFACE'12, but on the other hand its reaction time is slightly longer (1-2s) which can cause disturbing delays in the agent's actions; in particular, the agent should not speak simultaneously to the participants but would do so due to the introduced delay; this will be addressed in the future by consulting a low-delay voice activity detection feature when to decide if the agent can speak;
- the cheap microphones used were insufficient for the desired open scenario (agent audio rendered by loudspeakers), which created long laughter loops by the agent; high-quality directional microphones must be used in the future, or the audio of the agent should be rendered through headphones;
- the open-source speech synthesis system used with the Greta agent was not intelligible enough, which, in addition to bad timing of some reactions, lead some users to neglect the agent; a professional speech synthesis system will be used in the future to limit this problem;
- more voice/face/body features must be detected or improved; in parallel, the detected features should be synthesised by the virtual character;
- analysis of mimicry during human-human interaction is in progress on the data corpora recorded in the framework of the EU Project ILHAIRE; results will contribute to improved human-virtual character interaction.

**Acknowledgments.** The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°270780.

## References

1. Bachorowski, J., Smoski, M.J., Owren, M.J.: Automatic discrimination between laughter and speech. *Journal of the Acoustical Society of America* 110, 1581–1597 (2001)
2. Becker-Asano, C., Ishiguro, H.: Laughter in social robotics - no laughing matter. In: *International Workshop on Social Intelligence Design (SID 2009)*, pp. 287–300 (2009)
3. Becker-Asano, C., Kanda, T., Ishi, C., Ishiguro, H.: How about laughter? perceived naturalness of two laughing humanoid robots. In: *3rd International Conference on Affective Computing & Intelligent Interaction, Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6 (2009)

4. Bernhardt, D., Robinson, P.: Detecting affect from non-stylised body motions. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) *ACII 2007*. LNCS, vol. 4738, pp. 59–70. Springer, Heidelberg (2007)
5. Bourgeois, P., Hess, U.: The impact of social context on mimicry. *Biological Psychology* 77(3), 343–352 (2008)
6. Brand, M.: Voice puppetry. In: *Proceedings of Conference on Computer Graphics and Interactive Techniques*, pp. 21–28 (1999)
7. Bregler, C., Covell, M., Slaney, M.: Video rewrite: Driving visual speech with audio. In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997*, pp. 353–360. ACM Press/Addison-Wesley Publishing Co., New York (1997)
8. Cai, R., Lu, L., Zhang, H., Cai, L.: Highlight sound effects detection in audio stream. In: *Proceedings of the 2003 IEEE International Conference on Multimedia and Expo (ICME)*. Baltimore, USA, pp. 37–40 (2003)
9. Castellano, G., Villalba, S.D., Camurri, A.: Recognising human emotions from body movement and gesture dynamics. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) *ACII 2007*. LNCS, vol. 4738, pp. 71–82. Springer, Heidelberg (2007)
10. Cohen, M.M., Massaro, D.W.: Modeling coarticulation in synthetic visual speech. In: *Models and Techniques in Computer Animation*, pp. 139–156. Springer (1993)
11. Cosker, D., Edge, J.: Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations. In: *Proceedings of Computer Animation and Social Agents (CASA 2009)*, pp. 21–24 (2009)
12. Deng, Z., Lewis, J., Neumann, U.: Synthesizing speech animation by learning compact speech co-articulation models. In: *Computer Graphics International 2005*, pp. 19–25 (2005)
13. DiLorenzo, P.C., Zordan, V.B., Sanders, B.L.: Laughing out loud: Control for modeling anatomically inspired laughter using audio. *ACM Transactions on Graphics (TOG)* 27(5), 125 (2008)
14. Ezzat, T., Geiger, G., Poggio, T.: Trainable videorealistic speech animation. In: *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 57–64 (2004)
15. Fukushima, S., Hashimoto, Y., Nozawa, T., Kajimoto, H.: Laugh enhancer using laugh track synchronized with the user’s laugh motion. In: *CHI 2010 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2010*, pp. 3613–3618. ACM, New York (2010)
16. Gilroy, S.W., Cavazza, M., Niranen, M., Andre, E., Vogt, T., Urbain, J., Benayoun, M., Seichter, H., Billinghamurst, M.: Pad-based multimodal affective fusion. In: *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009* (2009)
17. Gosling, S., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. *Journal of Research in Personality* 37(6), 504–528 (2003)
18. Hofmann, J., Platt, T., Urbain, J., Niewiadomski, R., Ruch, W.: Laughing avatar interaction evaluation form. Unpublished Research Instrument (2012)
19. Kennedy, L., Ellis, D.: Laughter detection in meetings. In: *NIST ICASSP 2004 Meeting Recognition Workshop*, pp. 118–121. Montreal (May 2004)
20. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey (2012)
21. Kleinsmith, A., Bianchi-Berthouze, N., Steed, A.: Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 41(4), 1027–1038 (2011)

22. Knox, M.T., Mirghafori, N.: Automatic laughter detection using neural networks. In: *Proceedings of Interspeech 2007*, pp. 2973–2976. Antwerp, Belgium (2007)
23. Kshirsagar, S., Magnenat-Thalmann, N.: Visyllable based speech animation. *Comput. Graph. Forum* 22(3), 632–640 (2003)
24. Lasarczyk, E., Trouvain, J.: Imitating conversational laughter with an articulatory speech synthesis. In: *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, pp. 43–48 (2007)
25. Leite, I., Castellano, G., Pereira, A., Martinho, C., Paiva, A.: Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world settings. In: *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 367–374. ACM (2012)
26. Liu, W., Yin, B., Jia, X., Kong, D.: Audio to visual signal mappings with hmm. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004* (2004)
27. Mancini, M., Glowinski, D., Massari, A.: Realtime expressive movement detection using the eyesweb xmi platform. In: Camurri, A., Costa, C. (eds.) *INTETAIN. LNICST*, vol. 78, pp. 221–222. Springer (2011)
28. Mancini, M., Hofmann, J., Platt, T., Volpe, G., Varni, G., Glowinski, D., Ruch, W., Camurri, A.: Towards automated full body detection of laughter driven by human expert annotation. In: *Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction, Affective Interaction in Natural Environments (AFFINE) Workshop*, Geneva, Switzerland, pp. 757–762 (2013)
29. Mancini, M., Varni, G., Glowinski, D., Volpe, G.: Computing and evaluating the body laughter index. *Human Behavior Understanding*, 90–98 (2012)
30. Meng, H., Kleinsmith, A., Bianchi-Berthouze, N.: Multi-score learning for affect recognition: The case of body postures. In: D’Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) *ACII 2011, Part I. LNCS*, vol. 6974, pp. 225–234. Springer, Heidelberg (2011)
31. Niewiadomski, R., Mancini, M., Baur, T., Varni, G., Griffin, H., Aung, M.: MMLI: Multimodal multiperson corpus of laughter in interaction. In: *Fourth Int. Workshop on Human Behavior Understanding, in Conjunction with ACM Multimedia 2013* (2013)
32. Niewiadomski, R., Pelachaud, C.: Towards multimodal expression of laughter. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) *IVA 2012. LNCS*, vol. 7502, pp. 231–244. Springer, Heidelberg (2012)
33. Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfeller, F., McKeown, G., Pietquin, O., Ruch, W.: Laugh-aware virtual agent and its impact on user amusement. In: *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems, AAMAS 2013. International Foundation for Autonomous Agents and Multiagent Systems*, Richland, SC, pp. 619–626 (2013)
34. Oura, K.: HMM-based speech synthesis system (HTS) (computer program webpage), <http://hts.sp.nitech.ac.jp/> (consulted on June 22, 2011)
35. Owens, M.D.: It’s all in the game: Gamification, games, and gambling. *Gaming Law Review and Economics* 16 (2012)
36. Petridis, S., Pantic, M.: Audiovisual discrimination between laughter and speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, pp. 5117–5120 (2008)

37. Petridis, S., Pantic, M.: Is this joke really funny? Judging the mirth by audiovisual laughter analysis. In: Proceedings of the IEEE International Conference on Multimedia and Expo, New York, USA, pp. 1444–1447 (2009)
38. Poe, E.A.: Maelzel's chess-player. In: Southern Literary Messenger, vol. 2, pp. 318–326 (1836)
39. Qu, B., Pammi, S., Niewiadomski, R., Chollet, G.: Estimation of FAPs and intensities of AUs based on real-time face tracking. In: Pucher, M., Cosker, D., Hofer, G., Berger, M., Smith, W. (eds.) The 3rd International Symposium on Facial Analysis and Animation. ACM (2012)
40. Ruch, W., Ekman, P.: The expressive pattern of laughter. In: Kaszniak, A. (ed.) Emotion, Quality and Consciousness, pp. 426–443. World Scientific Publishers, Tokyo (2001)
41. Ruch, W., Proyer, R.: Extending the study of gelotophobia: On gelotophiles and katagelasticians. *Humor-International Journal of Humor Research* 22(1/2), 183–212 (2009)
42. Ruf, T., Ernst, A., Küblbeck, C.: Face detection with the sophisticated high-speed object recognition engine (shore). In: Heuberger, A., Elst, G., Hanke, R. (eds.) *Microelectronic Systems*, pp. 243–252. Springer, Heidelberg (2011)
43. Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Trans. Interact. Intell. Syst.* 2(1), 4:1–4:31 (2012)
44. Sundaram, S., Narayanan, S.: Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America* 121, 527–535 (2007)
45. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: ICASSP, pp. 1315–1318 (2000)
46. Tokuda, K., Zen, H., Black, A.: An HMM-based speech synthesis system applied to English. In: Proceedings of the 2002 IEEE Speech Synthesis Workshop, Santa Monica, California, pp. 227–230 (2002)
47. Truong, K.P., van Leeuwen, D.A.: Automatic discrimination between laughter and speech. *Speech Communication* 49, 144–158 (2007)
48. Urbain, J., Çakmak, H., Dutoit, T.: Arousal-driven synthesis of laughter. Submitted to the IEEE Journal of Selected Topics in Signal Processing, Special Issue on Statistical Parametric Speech Synthesis (2014)
49. Urbain, J., Cakmak, H., Dutoit, T.: Development of hmm-based acoustic laughter synthesis. In: Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech, Dublin, Ireland, pp. 26–27 (2012)
50. Urbain, J., Cakmak, H., Dutoit, T.: Evaluation of hmm-based laughter synthesis. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada (2013)
51. Urbain, J., Dutoit, T.: A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In: Proceedings of the Fourth Bi-annual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII 2011), Memphis, Tennessee, pp. 397–406 (2011)
52. Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J.: Avlaughtercycle: Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation. *Journal on Multimodal User Interfaces* 4(1), 47–58 (2010); special Issue: eNTERFACE 2009



53. Urbain, J., Niewiadomski, R., Mancini, M., Griffin, H., Çakmak, H., Ach, L., Volpe, G.: Multimodal analysis of laughter for an interactive system. In: Mancas, M., d' Alessandro, N., Siebert, X., Gosselin, B., Valderrama, C., Dutoit, T. (eds.) *Intetain. LNICST*, vol. 124, pp. 183–192. Springer, Heidelberg (2013)
54. Wagner, J., Lingenfeller, F., Baur, T., Damian, I., Kistler, F., André, E.: The social signal interpretation (ssi) framework - multimodal signal processing and recognition in real-time. In: *Proceedings of the 21st ACM International Conference on Multimedia*, Barcelona, Spain, October 21-25 (2013)

# Tutoring Robots

## Multiparty Multimodal Social Dialogue with an Embodied Tutor

Samer Al Moubayed<sup>1</sup>, Jonas Beskow<sup>1</sup>, Bajibabu Bollepalli<sup>1</sup>,  
Ahmed Hussen-Abdelaziz<sup>5</sup>, Martin Johansson<sup>1</sup>, Maria Koutsombogera<sup>2</sup>,  
José David Lopes<sup>3</sup>, Jekaterina Novikova<sup>4</sup>, Catharine Oertel<sup>1</sup>, Gabriel Skantze<sup>1</sup>,  
Kalin Stefanov<sup>1</sup>, and Gül Varol<sup>6</sup>

<sup>1</sup> KTH Speech, Music and Hearing, Sweden

<sup>2</sup> Institute for Language and Speech Processing- “Athena” R.C., Greece

<sup>3</sup> Spoken Language Systems Laboratory, INESC ID Lisboa, Portugal

<sup>4</sup> Department of Computer Science, University of Bath, UK

<sup>5</sup> Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany

<sup>6</sup> Department of Computer Engineering, Boğaziçi University, Turkey

{sameram,beskow,bajibabu,vhmj,catha,skantze,kalins}@kth.se,  
ahmed.hussenabdelaziz@rub.de, mkouts@ilsp.athena-innovation.gr,  
zedavid@l2f.inesc-id.pt, j.novikova@bath.ac.uk,  
gul.varol@boun.edu.tr

**Abstract.** This project explores a novel experimental setup towards building spoken, multi-modally rich, and human-like multiparty tutoring agent. A setup is developed and a corpus is collected that targets the development of a dialogue system platform to explore verbal and nonverbal tutoring strategies in multiparty spoken interactions with embodied agents. The dialogue task is centered on two participants involved in a dialogue aiming to solve a card-ordering game. With the participants sits a tutor that helps the participants perform the task and organizes and balances their interaction. Different multimodal signals captured and auto-synchronized by different audio-visual capture technologies were coupled with manual annotations to build a situated model of the interaction based on the participants personalities, their temporally-changing state of attention, their conversational engagement and verbal dominance, and the way these are correlated with the verbal and visual feedback, turn-management, and conversation regulatory actions generated by the tutor. At the end of this chapter we discuss the potential areas of research and developments this work opens and some of the challenges that lie in the road ahead.

**Keywords:** Multiparty, Multimodal, Turn-taking, Tutor, Conversational Dominance, Non-verbal Signals, Visual Attention, Spoken Dialogue, Embodied Agent, Social Robot.

# 1 Introduction

Today, advanced, reliable and real-time capture devices and modeling techniques are maturing and becoming significantly more accessible to researchers. Along with that, new findings in human-human conversations shed more light on the importance of modeling all the available verbal and non-verbal actions in conversations (in addition to the stream of words) and on how these are required in order to build more human-like dialogue systems that can be used by avatars and robots to exhibit natural behaviors (e.g. [1,2]). With these developments, research has been moving towards analyzing multiparty, multimodal conversations with the aim of understanding and modeling the structure and strategies with which interlocutors regulate the interaction, and keep their conversations rich, fluent, and successful.

Building socially aware and affective spoken dialogue systems has the potential of not only providing a hands-free interface for information input and output, but perhaps even more importantly, in many applications, the ability of using speech to provide a human-like interface that can understand and communicate all the subtle non-verbal signals that accompany the stream of sounds and provide significant information about the state of the user and the interpretation of the users verbal actions. These signals become even more central in scenarios where affective and social skills are essential for the success of the interaction (such as learning, collaborative task solving, games, and commerce [3-5]). Although the challenges and potentials of such social and affective technology are far from explored and understood, thanks to the recent availability and robustness of capture devices (e.g. microphone arrays, depth sensors), modeling techniques (e.g. speech recognizers, face tracking, dialogue modeling), and flexible and human like synthesis devices (e.g. avatars and humanoid robots), several recent projects are targeting the potential of different applications and high-end effects of modeling social and affective spoken interactions (e.g. Collaborative task solving in [6]; Education in [7]; Child therapy in [8]).

One major obstacle in the face of exploring the effects of spoken social and affective behavior of artificial embodied entities lies in the multidisciplinary nature of these setups and in the limitations of the different technologies that they involve. For example, while these applications aim at stimulating natural, fluent and spontaneous spoken behavior from the users, yet Automatic Speech Recognition systems (ASRs) still suffer a very limited power in handling such spoken conversational utterances, acoustically and grammatically. Another important challenge is how to keep these setups noninvasive, without hindering the fluency and spontaneity of the interaction (avoiding the use of cables, headsets, gaze trackers that dictate very little movement space in order for them to robustly function, etc.).

In this project, we target the development of a relatively natural, spoken, spatially and socially aware embodied talking head paying special attention to the aforementioned criteria.

The experimental design in this project is targeted towards multiparty collaborative task-solving, a research application that we expect to be central to the use of these technologies in the future. Such an application area is also rich with non-verbal and conversational variables that go beyond the meaning of words the users are using, but

extends to measuring other variables that play an important role in the interaction strategies and regulatory actions the agent should take into account, such as attention and conversational dominance.

## 2 Overview: The Moon-Survival Multiparty Tutor

Our work attempts to address interactional skills required by an embodied dialogue system to control the interaction flow as well as to boost and balance the engagement of the participants in the task they are involved in, while at the same time mitigating dominant behavior and encouraging less involved interlocutors to equally participate in the interaction. The task and the setup chosen in this work are considered as first steps towards understanding the behavior of a conversational tutor in multiparty task solving setups, as an example of a setup that can be used for applications in group-collaboration and negotiations, an activity that is highly dependent on the affective, and social behavior of the interlocutors [3]. Another main criterion that is taken into account when developing this setup is the ability to move directly from the models learnt from the annotations and analysis of the corpus, into an implementation of multiparty multimodal dialogue system, using the robot head Furhat [9], and the newly developed IrisTK dialogue platform [10] both developed and utilized in multimodal multiparty embodied spoken dialogue systems.

The interaction setup in this work consisted of two users and one tutor sitting around a round table. The two users' task is to discuss and negotiate the importance of certain objects and arrive to a decision on ordering them in terms of priority. The task was based on a shortened version of a "NASA Exercise: Survival on the Moon". During this exercise participants have to imagine that they are members of a space crew originally scheduled to rendezvous with a mother ship on the lit surface of the moon. However, due to mechanical difficulties, their ship was forced to land at a spot some 200 miles from the rendezvous point. During reentry and landing, much of the equipment aboard was damaged and, since survival depends on reaching the mother ship, the most critical items available must be chosen for the 200-mile trip. Two participants were presented with six cards with the pictures of six items left intact and undamaged after landing, as shown in Figure 2.

The tutor's task was to present the game, control its flow, and guarantee a high and balanced level of involvement between the two users and a collaborative decision process regarding the importance of the cards.

The remaining of the book chapter describes the project design and implementation that was done during the eNTERFACE2013 Workshop, that utilizes the Moon Survival Setup, as well as the requirements and development of the different technologies required to build a completely autonomous embodied dialogue system that could play the role of the tutor in similar setups. We firstly present a corpus collection study of a human-human setup, the design decision taken to reflect some of the affective and higher level conversational features that are present in collaborative task-solving, and outline some preliminary analysis of the data. After that, we describe a dialogue system setup where the human-tutor is replaced with the Furhat embodied talking head.

We also discuss the limitations of the work done, and possibilities for future research in this young and challenging area.

### 3 Experimental Setup

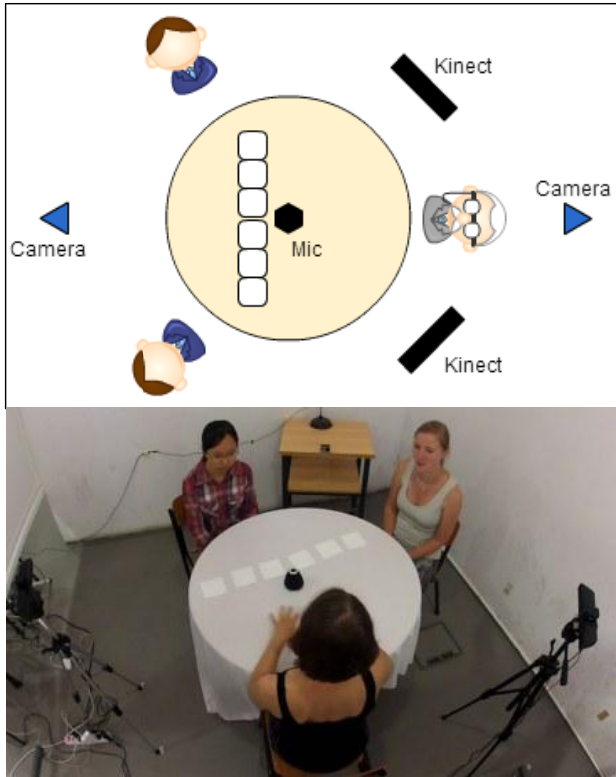
- *Physical setup*: Consists of a tutor and two participants, sitting at a round table, shaping an equilateral triangle.
- *Visual tracking*: Both subjects are tracked using two Kinect<sup>1</sup> sensors. The sensors are intended to capture the head-pose, facial expressions, and skeletal movement of both interlocutors. The Kinect sensors are placed at about 1.5 meter distance and outside the space of the interaction to limit the interference of the sensor on the interaction.
- *Auditory tracking*: Instead of using head-held close-range microphones (commonly used in dialogue recordings to limit the influence of overlapping speech); in this recording we took advantage of the Microcone<sup>TM2</sup> multichannel microphone array - by Dev-Audio. Microcone<sup>TM</sup> consists of 6 channel microphone (over 360 degrees) that provides high quality far-field speech input, along with activation values for the different microphones, allowing for the detection of multiple speakers, and hence overlaps and speakers locations. Microcone<sup>TM</sup> was designed for automatic annotation of roundtable meetings and it provides a measure of the microphone activity at 20fps. This means that the device is able to infer the speakers location (even in cases of overlap), and would provide raw audio signal for all six microphones with a reliable beam-forming and noise suppression. The choice of a table-top microphone array over a headset is made for two reasons: if people eventually address a dialogue system using a headset, the speech might be highly different from addressing a human (e.g. in terms of loudness). Also, avoiding cables and invasive attachments to the users might limit the influence of the experimental setup on the naturalness of the behavior, and the interaction might reflect patterns similar to that of a non-rigged natural interaction.
- Two high definition video cameras were used to record the setup and the interlocutors from two different angles, for future use and for annotation purposes. These cameras were used for (a) capturing tutor's behavior and (b) the entire scene.
- Six rectangular cardboard cards are designed and used as part of the game. The design of the cards was strategic in that it was made also to provide the dialogue system with context, and lower the demand for very robust speech recognition to infer the context of the game (e.g. When using vision-based card tracking, a system can infer the card under discussion, and provide spoken content related to it, without the need to understand what the users are saying).

Figure 1 shows a sketch of the physical setup to the left, and a snapshot of it with real users to the right.

---

<sup>1</sup> <http://www.microsoft.com/en-us/kinectforwindows/>

<sup>2</sup> <http://www.dev-audio.com/products/microcone/>



**Fig. 1.** Sketch (top) and a photo (bottom) of the experimental setup



**Fig. 2.** Six cards with pictures of objects for the task

The participants were asked to discuss each of the six cards and rank them in terms of importance. The motivation behind this task was to socially interact, exchange information, and collaboratively find a solution for the set problem. The members of each team collaborate together to solve the task. The tutor leads, coordinates, and gives comments to the team members while solving the tasks using several real-time strategies, e.g. using a *neutral* or *active* tutoring approach.

The described survival exercise was used for two reasons:

1. Groups first had to make descriptive judgments regarding the "value" of each item; and then they had to make judgments about the relative value of each item to their survival chances. Thus, both members of a group had to collaboratively participate in the conversation.
2. An important issue was the ability to compare participants' results with a right answer for the task, which was published by *the Crew Equipment Research Unit at NASA*. Group effectiveness was measured as a simple inverse function of the unit weighted sum of the absolute differences between the ranks assigned and the correct ranks. As we used a simplified version of the task, the overall performance of each group was evaluated in terms of time of task completion, in addition to the effectiveness.

## 4 A Corpus of Multiparty Tutoring Behavior

Recently, the research community has witnessed the birth of several large efforts towards the creation of large-scale multimodal corpora [11-15], promising that modeling verbal and visual signals in dialogue will not only advance the understanding of human-human language exchange, but also allow for the development of more intelligent and aware dialogue systems to be used by digital entities (such as ECAs and robots). However, there exist a multitude of design decisions (such as the dialogue task, the spatial setup, the captured signals) that limit the ability to easily move from human-human dialogues to human-machine dialogues. For example, dialogue tasks that heavily depend on the semantics in the speech signals (the content of the spoken interaction) will demand high requirements on speech understanding systems that can deal with conversational speech – a technology that has not yet been matured.

### 4.1 Users and User Setup

Before the recording session, participants were asked to complete a Big Five personality test [16]. The Big Five personality traits are five broad domains that are used to describe human personality. The Big Five factors are 1) openness to experience, 2) conscientiousness, 3) extraversion, 4) agreeableness, and 5) neuroticism. Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety. Conscientiousness shows a tendency for self-discipline. Extraversion is the personality trait of seeking fulfillment from sources outside the self or in community. High scorers in extraversion section tend to be very social while low scorers prefer to work on their projects alone. Agreeableness reflects a tendency to cooperate and adjust behavior to suit others. Neuroticism is the personality trait of being emotional and refers to a degree of emotional stability. We used the personality test because research indicates that personality traits and variables like self-efficacy self-esteem, locus of control, emotional stability, extraversion, conscientiousness, positive affectivity, negative affectivity, optimism, proactive personality [17], highly impact human work

results and performance. In addition, such factors as low neuroticism in combination with high extraversion characterize work engagement [18].

For the experiment the groups were formed according to participants' personality-test results, so that one of two team members scored high on extraversion and the other one scored low. The average difference between participants on the extraversion dimension was 28 points.

The human tutor was instructed to behave in a *neutral* way with four out of eight groups. A *neutral* tutor had to deliver material in a clear and concise manner so that participants could understand what they were required to do. However, a neutral tutor didn't need to make his/her communication either interesting or enjoyable. A neutral tutor had to answer all the students' questions, coordinate their activity and explain what to do next, but a neutral tutor didn't have to try to engage students and motivate them. A neutral tutor did not need to be friendly, supportive or welcoming. For the latter four groups, the tutor was asked to behave in a way that best represents the approach of an *active* tutor. An *active* tutor had to be dedicated to a student's success, had to deliver material in an interesting manner so that students could enjoy it. An active tutor had to be supportive, friendly and welcoming and always providing a positive feedback to the student.

Eight recording sessions were performed; each session resulted approximately in 10-15 minutes conversation. Afterwards the participants were asked to fill in a Tutor Assessment Questionnaire. The assessment questionnaire was based on the User Experience Questionnaire UEQ [19] and it consists of twenty four pairs of contrasting characteristics that may apply to the tutor. The numbers between the characteristics represent gradations between the opposites. A seven-step Likert scale is used for gradation in order to reduce the well-known central tendency bias for such types of items. Please refer to Appendix A for the full questionnaire.

Twenty four characteristics were organized into groups, suggested by [19]. These groups were Attractiveness (examples for items: pleasant, enjoyable), Perspicuity (clear, easy to understand), Efficiency (fast, organized) and Dependability (supportive, meets expectations). We also had an additional group called Tutoring with items specific to the tutoring approach, e.g. motivating, holding the attention, giving feedback on the work's quality.

Validity of the used questionnaire was tested by measuring the consistence of each group, as proposed by the original UEQ [19]. The Cronbachs Alpha-Coefficient [20], for Attractiveness, Efficiency, and Perspicuity and Tutoring groups was between 0.72 and 0.95. There is no generally accepted rule on how big the value of the coefficient should be, however many authors assume that a scale should show an alpha value  $>0.7$  to be considered as sufficiently consistent. Based on the high value of the Cronbachs Alpha-Coefficient we assume that the given groups of items in the questionnaire were consistent and that our participants in the given context interpreted the items in an expected way.

The data collected after 8 sessions (four groups with an active human tutor and four – with a neutral one) with 16 subjects an average assessment results were higher for an active tutor in all the assessment sections – attractiveness, perspicuity, efficiency, dependability and tutoring, as shown in Figure 3.



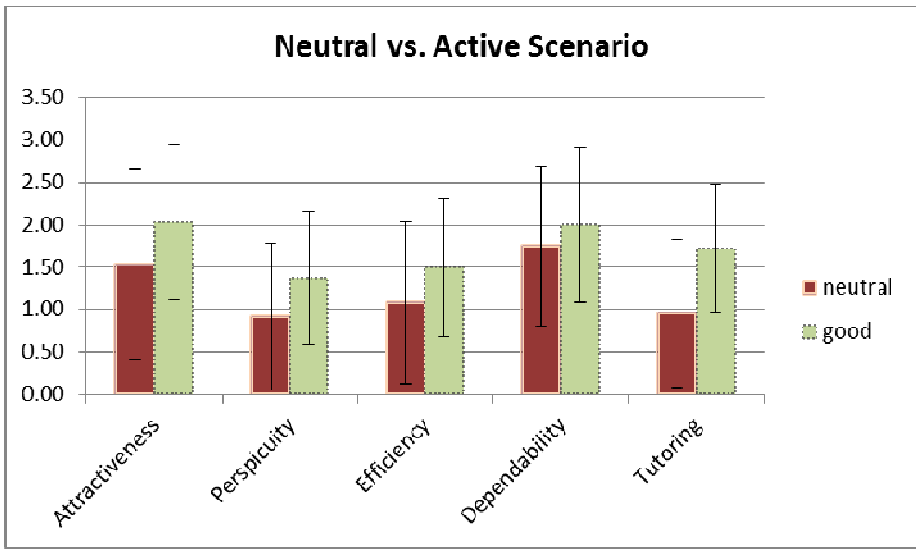


Fig. 3. Differences in tutor’s assessment for an active and a neutral tutor

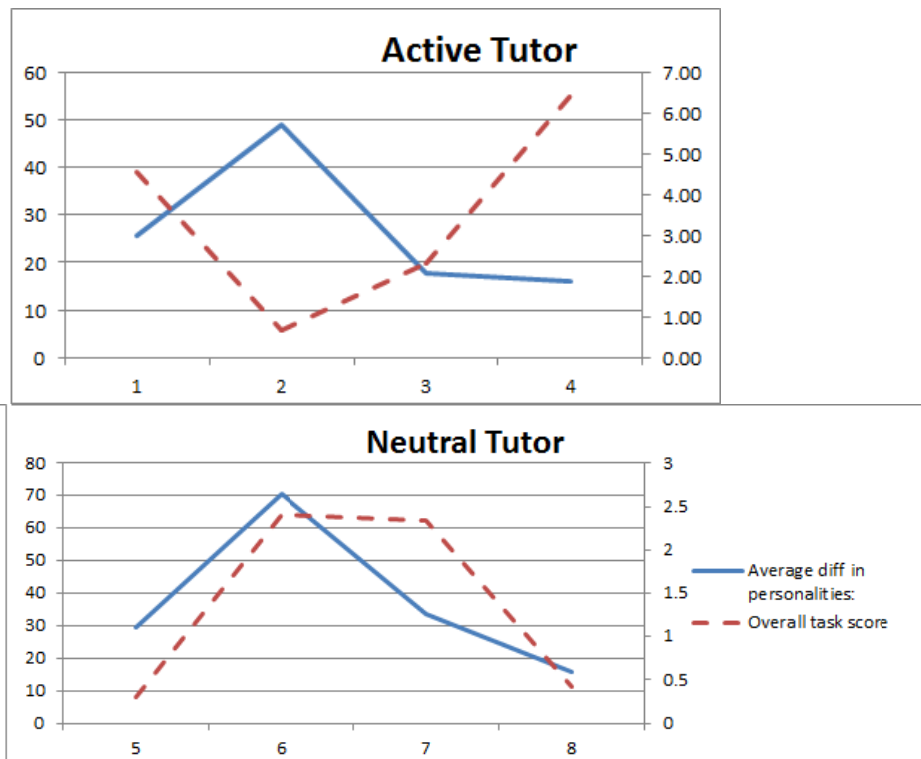


Fig. 4. The relation between differences in personalities and an overall task score in the case of active and neutral tutoring approach

Although differences between active and neutral tutoring approaches didn't influence significant differences in tutor's assessment results, these two different tutoring approaches caused different overall performance for the groups of subjects with different personalities. Figure 4 shows that in case of active tutoring, when the average personality difference between group members is high, the overall task score is low, which means that the performance of that group is better. On the contrary, if the average personality difference between group members is low, the overall task score of that group is high, which means that the performance of that group was worse. In case of neutral tutoring, however, there was no such an inverse dependency between personality differences and overall task score: the more different are the group's members according to their extraversion the lower is their task performance.

Thus, according to the presented data we can argue that active tutoring eliminates differences in personalities and helps groups with higher personality gaps achieve better results in a collaborative task.

The recorded corpus was named the *Tutorbot Corpus*.

## 4.2 Corpus Description

The eINTERFACE'13 *Tutorbot* corpus is of approximately 82 minutes overall duration and it consists of 8 sessions between a human tutor and two human participants. As described in the previous section, it includes equal samples of experimental conditions, i.e. 4 sessions of active tutoring interactional behavior and 4 of neutral. The tutor was the same subject in all sessions and was trained to express and prompt the appropriate conversational behavior according to each experimental condition. The participants, different in each session, were not informed about the task nor the goal of the experiment (participants of other projects in the eINTERFACE workshop). Depending on the availability of subjects, pairing subjects aimed at maximizing coverage in terms of personality traits (with a focus on the *extraversion* dimension) and gender. Details on the corpus are shown in Table 1.

**Table 1.** The Tutorbot corpus description

Session	Participants (Male/Female)	Duration	Tutor scenario	Extraversion diff.	Task Score
1	M-M	13.28	Active	37	4.56
2	M-M	15.44	Active	44	0.66
3	F-F	07.08	Active	5	2.30
4	M-M	10.43	Active	21	6.45
5	M-F	07.15	Neutral	9	0.30
6	M-M	09.39	Neutral	74	2.40
7	M-M	08.07	Neutral	13	2.34
8	M-F	10.05	Neutral	21	0.42

### 4.3 Annotation Process

The data collection was manually annotated with regards to the conversational behavior of the tutor. Since the goal is analyze multimodal strategies employed by the tutor to manage the conversation, the annotation was focused on describing the form and functions of the related verbal and non-verbal signals employed.

Multimodal interaction in both its two-party and multi-party dimensions is substantially related to the functions of feedback and turn management. Turn-taking mechanism has been thoroughly studied in terms of modeling the organization of turns in conversation [21], non-verbal cues such as gaze and gesture regulating turn taking in interaction [22], as well as the relationship between turn-taking and attention [23]. Multiparty turn-taking in dialog systems has been addressed with regards to the development of computational frameworks able to handle multiparty floor coordination, continuations, etc. [24]. The above studies focus on ways in which turn allocation is performed, rules that apply in transition-relevant places as well as aspects of collaborative and non-collaborative interactions such as interruptions and overlap resolution devices, both from verbal and nonverbal perspectives.

Communicative feedback gives evidence of the collaborative nature of dialogue while the participants give verbal and non-verbal signs that they follow the flow of the discussion; they perceive, understand, agree or not with the message conveyed; they might express the willingness to take the turn or give support to the speakers to go on with their turn. Feedback has been addressed in its linguistic dimension through a robust theoretical framework [25] and from a multimodal point of view with an emphasis the investigation of the effect that a combination of cues (e.g. morphological categories, prosody, gaze) might have on the production of feedback [26]. Moreover, there are attempts to describe feedback in different social activities or other contexts or model it for purposes of behavior simulation [27].

### 4.4 Annotation Scheme

The annotation of the recorded video sessions was performed in ELAN<sup>3</sup> [28]. An annotation scheme was employed to cater for all the features that need to be represented for the task at hand. The scheme is heavily based on widely-used labeling sets used for annotating multimodal interaction [29, 30], and was tailored to the needs of the task. Specifically, the goal of the annotation was to account for multimodal behavior including verbal and nonverbal signals as well as conversational structures and functions expressed in a multimodal way. Signals that have a clear communicative function are included in the annotation scheme as follows:

**Speech Activity.** The tutor's speech was transcribed with the goal to export utterances that the robot would use to manage the interaction. A comparison of the transcriptions was also planned, to distinguish patterns of verbal content when referring to specific subtasks in the discussion, i.e. in introducing the task, giving hints,

---

<sup>3</sup> ELAN (<http://www.lat-mpi.eu/tools/elan/>).

instructing the participants to order cards, etc. as well as to discover substantial differences of verbal content in active and neutral tutor scenarios. This level includes also the transcription of verbal back-channeling (grunts such as “yeah”, “ehm”, “aha”) the tutor may express.

**Dialogue Acts.** The tutor’s speech activity was attributed a label of a dialogue act describing the communicative action which the tutor performs. The purpose is two-folded: (a) to identify dimensions of interaction that dialogue acts may address and (b) to functionally segment the dialogue. Since in this experimental setup the identification of the addressee is of primary importance, the information-seeking functions (i.e. questions) are categorized not in terms of question types (e.g. yes/no question, wh-question), but in terms of addressee: questions to *speaker*, *listener*, or *both participants*. A crucial part of this scenario is the cues that the tutor provides to help the participants elaborate on the cards description and their importance (i.e. *hint*). Introductory parts where the tutor asks the participant to perform an action or clarifications given throughout the discussion are labeled as *Instruction/Request*. Finally, the scheme caters for *answers* that the tutor gives to the participants or utterances of *agreement* or *disagreement* with them.

**Turn Management.** Values in this level describe the way the tutor regulates the interaction by taking, holding and assigning the turn. Again, the values apply to both verbal and non-verbal behavior of the tutor. Different values exist for normal transition of turns (*take*, *accept*, *complete*, *offer*), as well as for phenomena related to aberrations from the turn-taking rules, such as interruptions and overlapping talk (i.e. *grab*, *yield*, *hold*). A distinct value of *backchannel* is also included to differentiate backchannel cues from content utterances.

**Feedback.** Labels related to feedback are attributed horizontally to cover both functions of verbal and non-verbal attestations of feedback, i.e. either through back-channeling and expressing evaluations, or through head movements and facial expressions such as nodding and smiling. The set consists of labels describing whether the tutor gives or elicits continuation, perception and understanding, and whether he/she agrees or not with what the participants say.

A large part of the annotation scheme is related to the annotation of the non-verbal modalities. Since the goal of the annotation is to identify important features and patterns to be modeled in the robot, the modalities in question are restricted to descriptive and functional values of the head movements, facial expressions and facial gestures, cues that are considered of high importance to the regulation of the interaction as well as the expression of feedback. Each non-verbal signal of the ones listed below is first identified on the time axis and it is marked according to its form. Subsequently, the functions of each identified signal is marked, i.e. whether it has a feedback or a turn management purpose.

**General Facial Expression.** The tutor's facial expressions are indicative of his/her state of mind towards the speakers as well as of the level of perception of the discussion. *Smile* and *laugh* are employed to show agreement, encouragement and satisfaction, while *scowling* denotes doubt, disagreement or unpleasantness.

**Head Movement.** The form and the direction of the head movement are important for establishing feedback and turn regulating functions. For example, head *nodding* may have an acknowledgement function, by providing support to the speakers that their contribution has been perceived and that the conversation may proceed. Head *turn* is always linked with gaze to determine attention and speaker turn assignment. *Shaking* is a sign of disagreement or doubt, while *tilting* the head or moving it *forward* and *backward* may be signals reinforcing the tutor's message.

**Gaze.** The identification of gaze direction is of primary importance since it defines the addressee of the tutor, the goal of his/her attention and can be a clear indicator of turn assignment. The scheme distinguishes between attentive gaze of the tutor to the speakers on the *left* and on the *right* respectively. Such values may be attributed i.e. when the tutor gazes at the speaker to provide feedback, but also towards the listener in an attempt to elicit feedback or to offer the turn. Attentive gaze at the *objects* (cards) is also substantial, since it indicates the tutor follows the task process. Non-communicative gaze shifts can be labeled as *glances*.

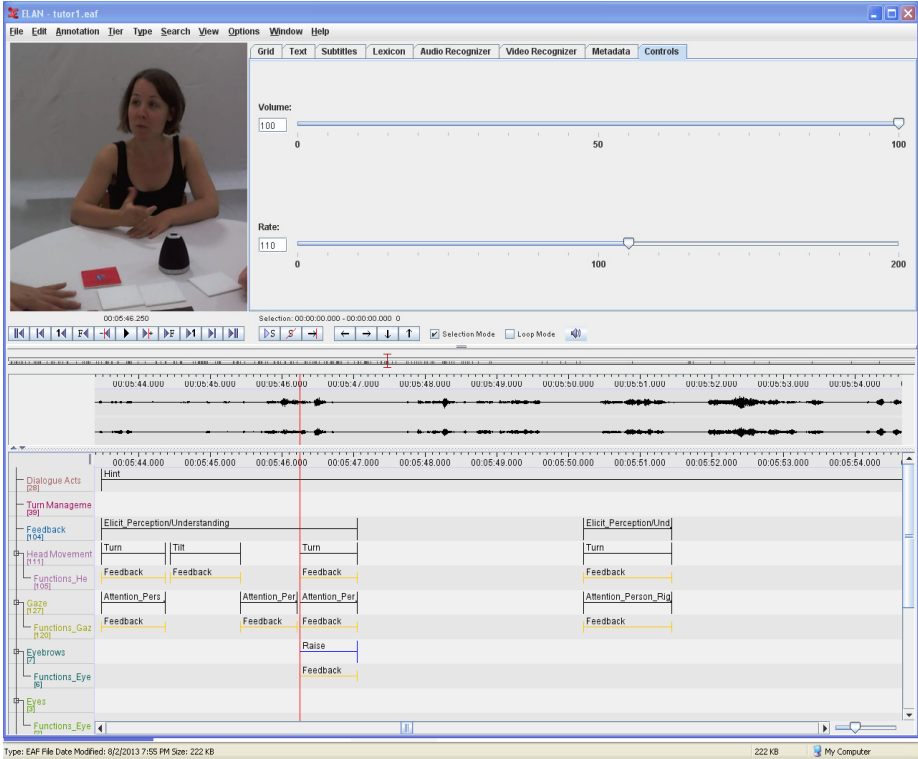
**Eyes.** Variations in eye openness may indicate surprise or enthusiasm (*wide open*), as well as contemplation, interest, attention or disagreement (*semi-closed*, *blink*).

**Eyebrows.** *Raising* eyebrows is often employed to show involvement, encouragement, attention and surprise, whereas *frowning* may denote doubt, disagreement or contemplation.

**Mouth.** An *open* mouth is annotated as a sign that the tutor is attempting to take the turn when a participant has the floor. A *closed* mouth with protruded lips may function as a feedback signal for agreement, together with head nodding.

**Cards.** Finally, a dedicated layer to the id of cards that are being discussed is included in the annotation scheme, so that the boundaries of each card object are clearly identified.

A tabular representation of the scheme may be found in Appendix B, Table 4. Figure 5 also shows a snapshot of the annotation program and process showing the tutor in the video.



**Fig. 5.** A screenshot of the annotation software. The video in the frame shows the tutor, along with the manual annotations over time.

#### 4.5 Data Analysis and Conversational Management Strategies

The data collected by all devices together with the manual annotation were analyzed to model the conversational management strategies employed in both conditions of *active* and *neutral* tutor. A set of different parameters was examined to attest the inter-relation of low-level signals such as voice activity, gaze, facial movements etc. and their timing with functions of turn management and feedback. Furthermore, differences in the tutor's dialogue acts and turn management behavior (in terms of frequency and different values employed) were investigated.

Our results indicate that turn management behavior conveys essential and richer information compared to the dialogue acts types used, i.e. the timing and the conversation managing action of what is said matters more than the actual content per se. For example, the number of turn offers as well as turn accepts is relatively higher in the *active* tutor condition than in the *neutral* one (42 vs. 8 and 33 vs. 13 respectively). We also hypothesized that: (a) the number of dialogue acts such as hints or instructions the tutor gives will be higher for the *active* tutor condition than for the *neutral* one and (b) the tutor will employ more turn management features in the *active* tutor condition than the *neutral* tutor condition. Concerning hypothesis a) we found a

difference in the number of hints between *active* and *neutral* tutor condition (29 vs. 27 hints) and concerning hypothesis b) we also found that the number of turn grabs is higher in the *active* than in the *neutral* tutor condition (13 vs. 8 turn grabs).

Overall, an important finding derived from the corpus verifying our hypotheses with regards to the experiment design is that almost all feature cases account for the expected interactional behavior in an *active* or a *neutral* tutor. A sample of statistics calculated on features is shown in Table 2.

**Table 2.** Statistics of features presented in the following order: *mean (standard deviation)*

<b>Feature</b>	<b>Active tutor</b>	<b>Neutral tutor</b>
Avg. time of all conversations	11.76(3.69)min	8.77(1.3)min
Avg. time on each card	1.29 (0.66)min	0.85 (0.49) min
No. of hints in all conversations	7.25(2.06)	6.5 (1.91)
No. of agreements	2 (1.82)	2.5(1.91)
No. of disagreements	2	1
No. of instruction/request	4.25 (0.95)	3.25 (0.5)
No. of turn grabs	3.25 (2.06)	2.25(2.06)
No. of turn offers	10.5 (8.38)	4.5 (5.74)

## 5 Building the Embodied Tutoring Agent

### 5.1 The Furhat Robot Head

The embodied agent used as the tutor in this project is the Furhat robot head [9]. Furhat was built to study and evaluate rich and multimodal models of situated spoken dialogue. Furhat is a robot head that consists of an animated face that is projected using a micro projector on a three dimensional physical mask that matches in design the animated face that is projected on it. The state of the art animation models used in Furhat produce synchronized articulatory movements in correspondence to output speech [31], and allow for highly accurate and realistic control of different facial movements. The head is also supported with a 3DOF neck for the control of its head-pose.

The solution to build a talking head using the technique used in Furhat is superior in that: 1) Using a three dimensional head allows for situated and multiparty interaction that is not possible to establish accurately with avatars projected on two dimensional surfaces, thanks to its ability to eliminate the so-called Mona Lisa gaze effect – an effect that results in a loss of the orientation of 2D portrait in physical space, resulting in that a viewer of a 2D face perceives the face rotated in the same angle no matter where the viewer is standing in relation to that portrait [32,33], and 2) The use of facial animation instead of other mechatronic solutions to build robot heads enables the use of highly advanced and natural dynamics that are not so easily possible with mechanical servos and artificial skin, thanks to the advanced in facial animation techniques [31]. Furhat, in addition to being a platform to implement models of spoken human-human interaction, has become a vehicle to facilitate research on human-robot interaction, such as studying the effects of gaze movements in situated



**Fig. 6.** Snapshots of Furhat<sup>4</sup> in close-ups

interaction [34], audio-visual intelligibility of physically three dimensional avatars [35], and effects of head-pose on accuracy of addressee selection [36]. Figure 6 shows some snapshots of the Furhat robot head.

## 5.2 The IrisTK Multimodal Multiparty Authoring Platform

To orchestrate the whole system, the IrisTK dialogue platform was used [10]. IrisTK is XML based dialogue platform that was designed for the quick prototyping and development of multimodal event-based dialogue systems. The framework is inspired by the notion of state-charts, developed in [37], and used in the UML modeling language. The state-chart model is an extension of finite-state machines (FSM), where the current state defines which effect events in the system will have. However, whereas events in an FSM simply trigger a transition to another state, state charts may allow events to also result in actions taking place. Another notable difference is that the state chart paradigm allows states to be hierarchically structured, which means that the system may be in several states at the same time, thus defining generic event handlers on one level and more specific event handlers in the sub-state the system is currently in. Also, the transition between states can be conditioned, depending on global and local variables, as well as event parameters. This relieves state charts from the problem of state and transition explosion that traditional FSMs typically leads to, when modeling more complex dialogue systems.

IrisTK is based on modeling the interaction of events, encoded as XML messages between different modules (a module can be a face tracker that transmits XML messages about the location of the face of a user). The design of module based systems is crucial in this system and in other multimodal dialogue tasks. Such systems are multidisciplinary in nature and researchers typically working on one of the technologies involved in such a system can be isolated from the details of the other technologies. This increases the need for the development of higher level, technology independent dialogue management that can allow for the communication of the different tools and technologies involved (Automatic Speech Recognizer - ASR, Text-To-Speech systems - TTS, Face

<sup>4</sup> For more info on Furhat, see <http://www.speech.kth.se/furhat>



Tracking, Facial Animation, and Source Localization). These technologies can be (and are, in this project) run on one or several machines.

IrisTK comes with several tools that support the communication of XML events in between programs, and over a network. This would allow the dialogue management to rely merely on XML events, while being able to be completely blind to the different programs that generate and consume these events. This also allows for the replacement of one or more program, or technology, without the need for any customization of the dialogue flow.

Relying on the principles of modular design and XML event communication protocol, we describe in the following the different sensory technologies that generated events, which in turn are consumed by the dialogue management flow (See Section 8 for design of the dialogue flow).

### 5.3 Modeling of Sensory Data

**Voice Activity Detection with the Microcone™.** In addition to providing audio, the Microcone™ also provides a stream of the current microphone activation status for each of its six audio channels. We used this stream to implement a voice activity detector (VAD) module for the system, by mapping microphone activation status transitions to start and end of speech.

The devised module generates a message when a subject starts speaking and when a subject stops speaking. Each message contains information about which subject triggered the message, and the amount of time passed since the previous transition, i.e. the length of silence before start of speech, or the length of the utterance at end of speech. Example XML events are presented below. Each event has a name, and a set of typed parameters. The following example shows two different events - a speech onset and speech offset time, providing parameters on which subjects is concerned with this event.

```
<event xmlns="iristk.event" name="sense.speech.start">
  <string name="location">Left</string>
  <float name="silence">4.5</float>
</event>
<event xmlns="iristk.event" name="sense.speech.end">
  <string name="location">Right</string>
  <float name="length">2.25</float>
</event>
```

As the experiment setup has the participants in fixed locations, the identity of a speaker can be mapped to an audio channel. The module keeps track of the current activation state of each channel in use, and creates events on activation state transitions, provided the state has been stable for a tunable time period. The tunable threshold allows for brief moments of silence in continuous utterances, and prevents short isolated sounds from triggering start of speech detection.

**Visual Tracking.** The behavior of the subjects was tracked partly using Kinect sensors (“Kinect for Windows”). In the experiment setup, one sensor was used for each subject and each sensor was placed in front of the subject it was monitoring. The intention was to have the subject facing the sensor, keeping the expected head pitch and yaw angles between  $\pm 30$  degrees.

The physical locations and orientations of the subjects’ heads, as well as parameters describing the facial expressions, were tracked using the Microsoft Face Tracking SDK (“Face Tracking”). Data from the head tracking allowed the system to have Furhat to direct its gaze at subjects, and to estimate the visual attention of the subjects. In addition to the tracking of heads, the poses of the subjects’ upper bodies were tracked as well, using the Kinect for Windows SDK skeleton tracking in seated mode (“Tracking Modes”). Skeleton data with ten tracked upper-body joints for each subject was collected for future use.

```
<event xmlns="iristk.event" name="sense.head">
  <string name="sensor">kinect_left</string>
  <string name="agent">Left</string>
  <float name="position.x">1.2</float>
  <float name="position.y">2.345</float>
  <float name="position.z">3.45678</float>
  <float name="rotation.x">1.2</float>
  <float name="rotation.y">2.345</float>
  <float name="rotation.z">3.45678</float>
  <float name="au.lipstretcher">-0.2</float>
</event>
```

**Visual Attention Estimation with Kinects.** We wanted to provide the system with information about the subjects’ visual focus of attention, which can be inferred from gaze direction. Complete gaze direction is, however, not available in the case of our setup, so an alternate method is required. One way of estimating the visual focus of attention without gaze is to use head pose information as a surrogate. This alternative was explored in [38], who in a round-table meeting scenario with four participants show an average accuracy of 88.7% for the estimation of focus of attention from head orientation alone. The contribution of the head orientation to the overall gaze was on an average 68.9%. A study in [39], expanded the meeting scenario with additional targets for visual attention, and found that the different targets for visual attention need to be well separated in order to achieve good estimation performance. In our experiment setup the potential targets for a subject’s attention is the tutor, the other subject, or the card on the table. The low target count, combined with the constraints of the experiment setup, suggest that we can use head poses as a good estimate for subjects’ visual focus of attention in our experiment, similar to a setup in [40].

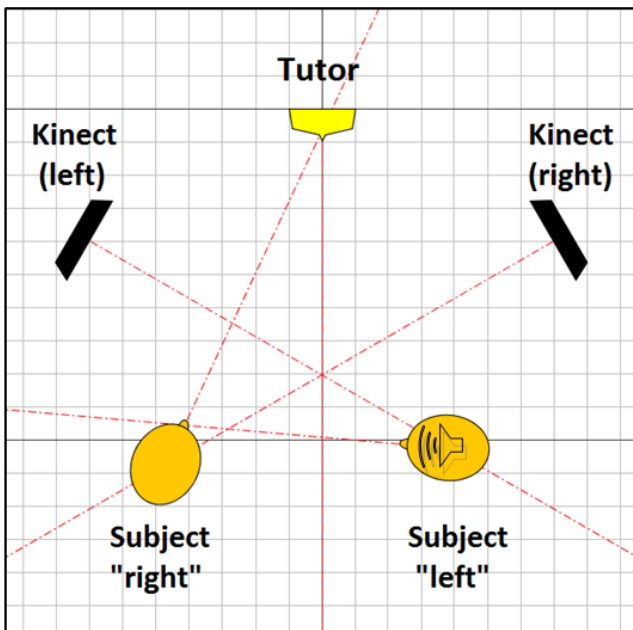
The devised visual attention module for our setup accepts head tracking data from the Kinect sensors and generates one message whenever the estimated visual attention target of a subject has changed. The message, illustrated below, contains information about which subject the message refers to, the direction of the subject’s head and the estimated target of the subject’s visual attention. Before the target is considered changed and a message is sent, the estimated target must be stable for a tunable period of time.

```

<event xmlns="iristk.event" name="sense.attention">
  <string name="agent">Left</string>
  <string name="direction">left</string>
  <string name="target">Right</string>
</event>

```

Since the physical setup ensures that each Kinect sensor is covering exactly one subject, the identity of detected heads could be derived from the sensor detecting it. The target of visual attention of a subject was estimated based on the pan and tilt of the subject's head. Each possible target was specified as a region defined by minimum and maximum angles to the target relative to the Kinect sensor located in front of the subject. The region boundaries for visual attention targets were calibrated manually for this experiment, a method we believe to be rather reliable for the current purpose due to the well-structured physical setup. We do, however, intend to improve the boundary definition by using more data driven methods for clustering in the future. Combining the microphone array and the visual attention classifier, the system can be informed about the speaker and the addressee at any given point in time. Figure 7 shows a visualization of the system in action, showing Speaker Left speaking to Subject Right, while subject right is looking at the tutor.



**Fig. 7.** A visualization of the perceived activity of the situated interaction. The figure shows the tracked horizontal head rotation (pan) of each subject, and the voice activity detection (highlighted by an image of a speaker). The figure shows that the subject (left) is currently speaking.

**Acoustical Prominence Detection.** In addition to the Speech Activity, and the Visual Attention modules, a prosodic analysis module was developed to estimate important segments in the users input using their prosodic features. Information about prosodic prominence can in principle enhance the speech recognizer by conditioning the syntactic parsing. In addition to that, it can also give important information to the agent, which in turn can show more contextually aware gestures in relevance to the users' prosodic contours [41]. In this work, we wanted Furhat to generate nonverbal gestures (as eyebrows raises) in response to prominent segments in the users speech, in order to show attentive behavior (such strategies have been shown to be functional in active listening experiments, c.f. [42]).

Acoustical prominence is perceived when a syllable or a word is emphasized so that it is perceptually salient [43]. Detecting the acoustical prominences can be very useful in Human-Robot interaction (HRI) scenarios. For example, the salience of the emphasized words uttered by human beings can be used as cues for triggering feedback signals generated by robots. The feedback signals can be acoustical by saying yeah or mmm, visual by raising eyebrow or smiling, or multimodal [44]. These feedback signals make the interaction between humans and robots more natural and increase humans' engagement in the conversation. Moreover, in multiparty dialogues, the frequency of the detected prominences can be used as a reasonable feature for detecting and balancing the conversational dominance of the dialogue participants.

In order to automatically detect acoustical prominences, some low level acoustic prosodic features should be first extracted. Features like F0, energy, and duration have shown a good success in automatic prominence detection [43]. Mapping these extracted prosodic features to the acoustical prominence, which is mostly defined based on linguistic and phonetic units, is conventionally done using annotated databases and supervised machine learning algorithms like neural networks (NN) [43], and hidden Markov models (HMM) [45]. In this work, however, we have used a modified version of the unsupervised statistical method applied in [46]. The main idea of this method is developed based on the prominence definition introduced in [47].

In [47], prominence is defined as the speech segment (syllable or word) that stands out of its environment. In order to realize this definition, we define a relatively short moving window in which the current speech segment (e.g. syllable) lies and another longer window for its preceding environment. We can simply detect the salience of the current speech segment by calculating a discrimination distance between the prosodic features that lie in it and those located in its preceding (environmental) window. Prominence is then detected if the distance is larger than a pre-defined threshold, which means that the current segment is salient and stands out of its environment.

The discrimination distance can be deterministic such as the Euclidean distance between the prosodic features' mean vector of the current and the environmental window, or probabilistic to take into account the uncertainty (covariance) of the feature vectors. A good candidate for the probabilistic discrimination distance is the Kullback-Leibler (KL) divergence [48]. By assuming that the prosodic feature vectors in the current local and the past global window are modeled by Gaussian distributions, the KL divergence can be computed via [49]:

$$D_{KL}(N_0||N_1) = \frac{1}{2} \left( \text{tr}(\Sigma_1^{-1}\Sigma_0)(\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k - \log \left( \frac{\det(\Sigma_0)}{\det(\Sigma_1)} \right) \right) \quad (1)$$

where  $\mu_0, \mu_1, \Sigma_0$  and  $\Sigma_1$  are the mean feature vectors and the covariance matrices of the current local window and the past global window, respectively. In (1.1),  $k$  is the feature vector dimension.

However, one problem of applying the KL divergence here is the difference in the estimation reliability of the global and the local window parameters. This difference arises due to the convention of choosing the local window length shorter than that of the global window. For that reason and the fact that the KL-divergence is non-symmetric, we have used instead a modified version of the  $T^2$  Hotelling distance:

$$D_H(N_0||N_1) = \frac{L_0 L_1}{L_0 + L_1} ((\mu_1 - \mu_0)^T W \Sigma_{0 \cup 1}^{-1} (\mu_1 - \mu_0)), \quad (2)$$

where  $L_0$  and  $L_1$  are the length of the local and the global window and  $\Sigma_{0 \cup 1}$  is the covariance matrix of the union of the samples of the local and the global windows. The main function of the added weight matrix  $W$  here is to give prosodic features different importance. However, if all the prosodic features are of the same importance then the weight matrix  $W$  will be the identity matrix  $I$  and the  $T^2$  Hotelling distance in (1.2) reduces to its standard form in [50].

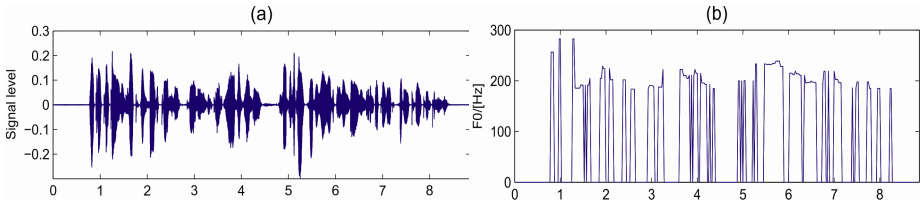
#### *Implementation aspects.*

The prosodic features used in this experiment are the fundamental frequency  $F_0$  and the energy  $E$ . To extract these features, we have implemented a real-time multi-channel prosodic feature extractor module. This module extracts the short time energy in dB via

$$E(t) = 10 \log \left( \frac{1}{K} \sum_{k=0}^K x^2(k, t) \right) \quad (3)$$

where the  $x(k, t)$  is the  $k^{\text{th}}$  sample of the  $t^{\text{th}}$  frame and  $K$  is the frame length. The  $F_0$  in this module is extracted according to the pitch tracking algorithm YIN [51] (e.g. see Figure 8). The prosodic features are extracted from short time frames of length 50 ms with 50% overlap between consecutive frames. In order to compensate the outliers, the output of the pitch tracker is applied to a median filter of length three.

In [52], it has been shown that the average duration of vowels in heavily stressed syllables is between 126 and 172 ms. Thus, the length of the local window has been tuned in this range so that the performance of the prominence detector is optimized. The length of the global window that models the environment of the current acoustic event is chosen to be seven times larger than the length of the local window length. The feature vectors used to calculate the  $T^2$  Hotelling distance in (2) are the feature vectors extracted only from voiced speech.



**Fig. 8.** (a) Exemplary input signal to the real time F0 tracker. (b) The estimated F0 using the YIN algorithm.

**Visual Tracking of Game Cards.** To allow the dialogue system to infer the status of the game, without the dependency on interpreting spoken content from the users, the game design employed six cardboard cards on each one of which an object was shown. Since the design of the game and the setup was not mandated by technical limitations, this allowed for flexibility to design and color the cards to maximize the accuracy of a card tracking system that is not sensitive to lighting changes. The design of the table and the game was also flexible, thus the game was designed so that subjects would place the cards in certain dedicated spots and with a certain orientation.

Detection, recognition and tracking of an arbitrary object in video stream are inherently difficult tasks. Most of the problems stem from the fact that light conditions change over time. Furthermore, abrupt motion, changes in shape and appearance and occlusions make the tasks more challenging. There is tremendous amount of research targeted at tackling these problems and many algorithms have been proposed in the literature [53].

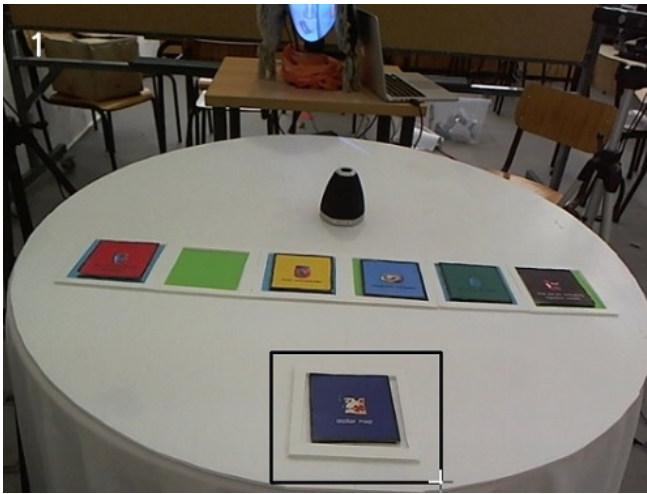
The main requirement for the developed system is to have real-time response. There are computationally feasible methods, which can compete with the more complex ones, given a set of assumptions [53]. Since we can design the game and the flow, we can safely assume that light condition will not significantly change during the game and the shape and the appearance of the tracked objects is constant.

The dialogue system was conditioned by input from the card tracking about the timing and an identity of a new card (whenever users flipped a new card to discuss it). Another requirement from the card tracking system was that whenever the users flip all the cards and put them in order, the card tracking system needs to inform the dialogue system that a new order has been established (the tracking system should also inform the dialogue system whenever a new order is in place – this could happen if the users discuss further the cards and change their agreement).

The cards are designed so that each one has a distinct color. This enables the system to differentiate the cards by comparing their color histograms. This type of comparison is convenient because the color histogram does not change significantly with translation and rotation.

The card tracking system accepted a video stream from a video camera that is directed towards the table (Figure 9). The system allowed the experiment conductor to initialize the templates of the cards whenever needed. This is assumed to be important

by the beginning of each interaction, as lighting changes might happen over large periods of time. In order to initialize the system, the user is required to define two regions of interest in the video. The first one is the region at the bottom of the table where the card under discussion resides (users were asked by the tutor agent to flip open a card and place it in the dedicated spot before they start discussing it). The second is the region in the middle of the table where all cards reside – this region of interest is used to track the order of all the cards whenever the cards are all flipped open at the end of the discussion. Figure 9 shows an image from the video stream of the camera used for the card tracking system, and illustrates a selection of the active card region of interest.



**Fig. 9.** Active card ROI

In order to recognize the card at the bottom of the table, first a contour detection is performed on the whole region of interest. All contours are then approximated with polygons and then each polygon is described by its bounding box. After filtering the resulting bounding boxes through predefined threshold (removing small false detections) the smallest bounding box is selected as the target object. The histogram of the crop of the target is calculated and the correlation between the target histogram and all template histograms is calculated. The closest template is chosen as the recognition result. If the recognition does not change for a predefined number of frames, the system broadcasts this decision to the dialogue manager.

After all the cards are discussed, the participants need to agree on order of relevance. This is done following the same algorithm used to track an active card. The cards are then sorted with respect to the top-left corner coordinates of their bounding box. Thus, we obtain the relevance information for each of them (e.g. the left most being the most important). Figure 10 illustrates real-time tracking and the pipeline of the implemented system.

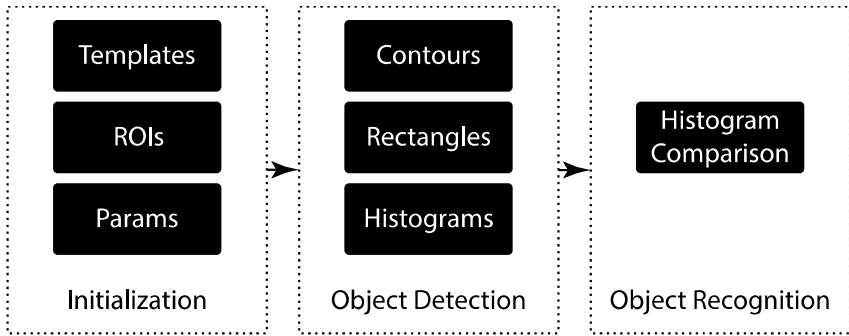


Fig. 10. Card tracking pipeline

## 6 Dialogue System

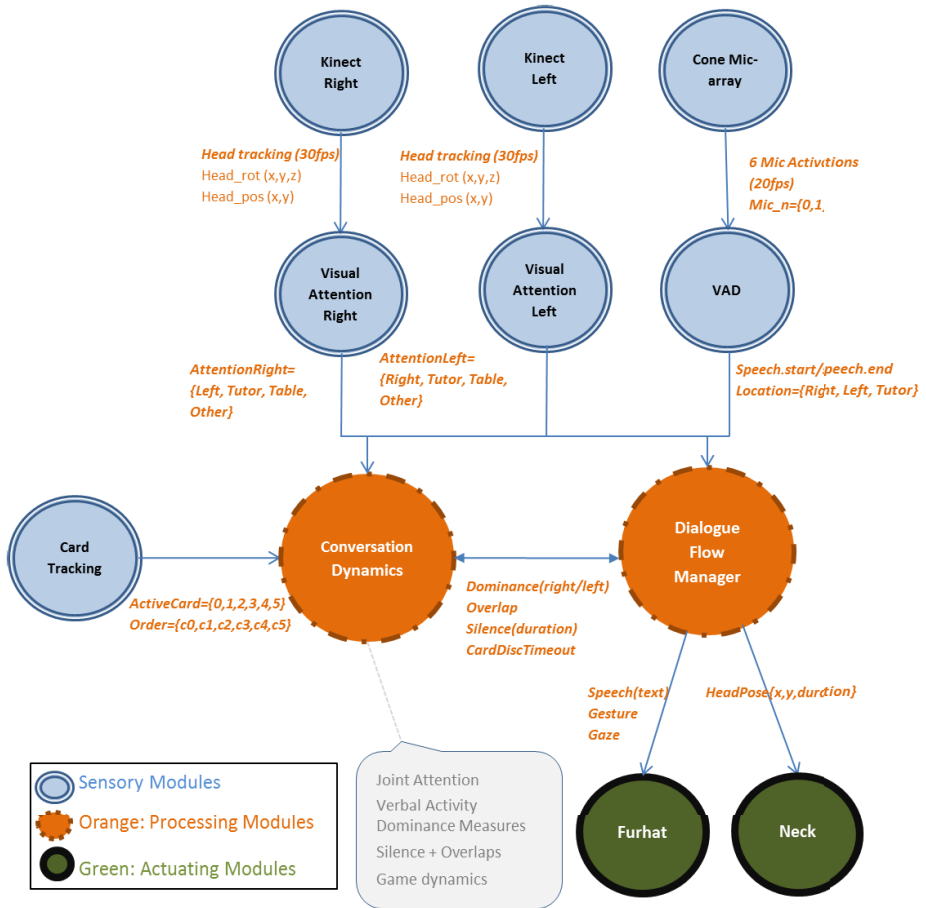
In this project, the IrisTK framework [10] for multimodal spoken dialog system is used to author the dialogue system, and in turn to control the Furhat robot head. This framework allows the incorporation of a series of modules and facilitates a standardized event-based communication between them. These events can represent input data (Sense), something that the system should do (Action) or feedback-loop sensation about a certain action to infer the physical state of the system (Monitor). The system architecture was designed to accommodate different setups according to the needs.

The sensory modules collect information from the environment:

- The card-tracking module is responsible for two different tasks: track the card under discussion and, when the discussion ends, provides the system with the card order. The details of how the card tracking is performed are detailed in the previous section.
- The two Kinects which track the orientation of the heads in order to infer visual attention. With the head position, the system can be informed about the location of the speakers (which would help the system establish mutual gaze when needed), and with the orientation, the system can track where the speakers are focusing their attention, and who they are addressing (for details about the visual attention module, please refer to Section 5.3)
- Finally, to combine the visual attention with the verbal activity, the setup employs one microphone array (Microcone™), composed of six microphones positioned around a circle covering in total 360 degrees. The microphones are used to perform Voice Activity Detection (VAD) for each of the participants in the dialogue.

The actuating modules perform the visible actions. The Furhat module is responsible for managing the agent's face, gaze, gestures and speech synthesis tasks. The Neck module performs head movements mainly by directing attention to the speakers, using input from the Kinect face tracking modules. Figure 11 shows a flow chart of the main modules of the dialogue system.





**Fig. 11.** An overall chart view, showing the flow of information in the system. Each circle represents an independent Module that communicates with other modules using events encoded as XML messages. Sensory modules are mainly responsible for providing input events to the system. Processing modules are responsible for modeling the dialogue using sensory events and internal state definitions. Actuating modules are responsible for activating the robot head.

### 6.1 Conversation Dynamics

Conversation Dynamics is a key module for Furhat’s tutoring task since it computes figures in real time (updated every 1ms) that relate to conversational properties of the interaction. Such events can be looked at as the main drive that will decide when the tutor should intervene in the dialogue. The principle behind supporting the dialogue manager with a “conversational dynamics” module is to build an up-to-date model of the interaction. Such model allows the dialogue manager access to high level states of the interaction, instead of calculating them as part of its dialogue state design. This will remove the need on the dialogue system to contain dedicated States for each and

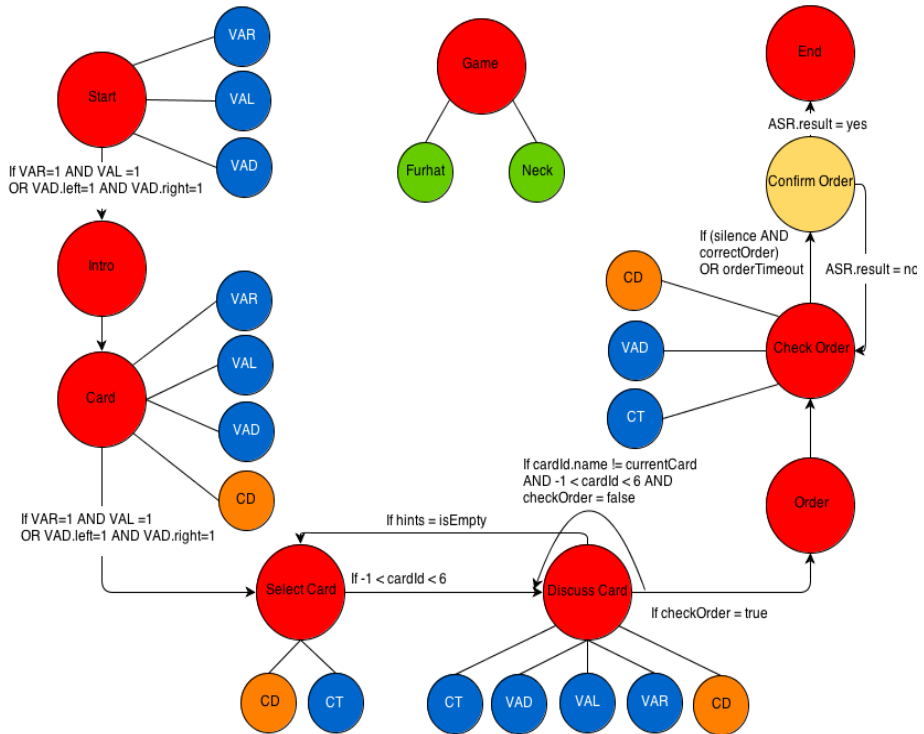
every possible combination of sensory input and context. Since the task of the system is to infer higher level conversational parameters and act on them (such as dominance, low levels of engagement), the conversational dynamics is responsible for containing variables about the verbal activity of each participant, their current visual target, and other long-term parameters, such as the percentage of silence a certain user has been in since the beginning of the dialogue. This allows the dialogue manager to access these parameters on demand. The conversational dynamics module is also responsible for firing events related to the interaction between users, for example, if users are silent for more than a specific threshold, the conversational dynamics module can send an event called “low engagement”, which the dialogue flow in turn can respond to by taking the initiative and directing a question to one of the participants.

- Verbal activity is computed for both speakers. For each of them a vector with the frames where each of them spoke in the last 200ms is computed, based on the Voice Activity Detection performed by the microphone array. These vectors are used to compute the dominance that relies upon the difference in the verbal activity between the two speakers over a longer period of time. Measuring dominance is a research topic by itself and has received considerable attention, where most work has targeted the offline annotation of meeting corpora. In [54] for example, Support Vector Machines (SVMs) were used for a posteriori classification of dominance in meetings. In our scenario, we needed a real-time dominance classifier. The solution adopted was a rule-based decision, using a threshold on the difference of verbal activity between the two speakers inspired by the analysis of the recorded corpus, using the timing of tutor interruption and turn management as thresholds. If the difference in the verbal activity is above that threshold, a dominance event will be generated. The verbal activity values are reset once the card under discussion changes.
- Conversation dynamics also computes the silence information for each of the speakers separately and joint silences for all participants (including Furhat). The period since speakers started speaking is also computed for each of them. The combination of speaking times results in the overlap speech period. Both are computed since the conversation started and since the card under discussion changed.
- This module also tracks the duration of the current card discussion and total discussion times. If these reach the thresholds set, events are generated to make the system suggest a change in the card under discussion or, in case of ordering, to suggest the end of the discussion.

## 6.2 Flow Description

The dialogue manager is specified using a state chart-based framework defining the flow of the interaction (IrisTK flow [10]). In our experiment two different flows were created, one for the *Neutral* tutor and another for the *Active* tutor. These tutors try to map the characteristics revealed by the different tutor behaviors in the sessions with the human tutor. The complete diagram flow is shown in Figure 12. The difference between the tutors is not the flow itself, but the way the states are implemented.

The first state defined in the flow is the Game state, a general state. All the other states in the flow will extend the specifications of this state, which means that the behaviors specified within this state would be available in every state that extends this one. These behaviors correspond to actions that the head must perform. The following bullets explain the structure of the dialogue system in terms of the states it occupies over time, and in relevance to the flow of the dialogue.



**Fig. 12.** Flow chart of the dialogue system states. VAD: Voice Activity Detection. VAL: Visual Attention of Left speaker. VAR: Visual Attention of Right speaker. CD: Conversational Dynamics. CT: Card Tracking input.

- The “Start” state is the initial state in the dialogue. In this state, the system greets the users for the first time and waits until both of them are detected to move on to the next state. This detection is performed either using visual cues (collected from Kinect) or audio cues (collected from the microphone array). If only one of the users is detected, the system informs her/him that they should wait for the other user to be detected in order start the discussion. Once both users are detected the flow goes to “Intro” state.
- In the “Intro” state, an explanation of the moon survival task is given to the user. After this explanation, the dialog proceeds to the “Card” state.

- The “Card” is just a step to check that the users are ready to play the game. The system checks that they are ready by detecting them using voice activity detection and visual attention state of both of them, and then moves the flow to the “Select Card” state. If they remain silent above the threshold silence time, than the system prompts a sentence to make the users speak and move to the “Select Card”.
- The “Select Card” state waits for a sense.card event generated from the card tracking module. The event contains the card id of the identified object. Until the id is valid or the users keep the silence the system is going to push them to select a valid card.
- If the valid id is detected the “Discuss Card” state is activated. In this state, the system manages the card discussion. The discussion management is different between the two types of tutor. In both of them silent periods are tracked.
- If the users have been silent for a specific time (a threshold is reached), the system is going to evaluate the user’s attention. In the human tutor dialogs, if both users were looking to the tutor and one of them asked something to the system, they are both waiting for an answer. The tutor should address both speakers and use one of the hints transcribed for this type of behavior. The hints for the objects are loaded as a stack. Once the hint is used it is popped out of the stack. It might occur that the hints for the object under discussion were all popped from the stack. In that case the system will encourage the users to pick another card.
- If only one of the users is looking at the tutor, and she/he was the last to talk, the tutor only addressed this user when answering. The same behavior was implemented, having the tutor answering towards the last user who talked and using one of the hints transcribed in the corpus for the object under discussion.
- The system is also measuring the total silence time. If the threshold is reached, the system should use one of the prompts that were used by the human tutor whenever there were long silences. These prompts should encourage the users to continue the conversation.
- There is also a timeout and minimum time for a card discussion. When the timeout is reached the system suggests the users to flip over a new card. If card event (change of card) is detected before the minimum discussion time is reached the system gives a hint about the card that was being discussed before the card event was detected, in order to make the users continue the argument about the card that they have decided to change. These thresholds were set based on the data collected in the human tutor corpus. Thus different thresholds were used for the Neutral and Active tutor flows.
- Another difference between the two configurations is how they deal with a dominance event generated by the Conversation Analysis. The Active tutor grabs the turn from the dominance speaker whenever the Conversation Dynamics module has generated the dominance event, whereas the Neutral tutor does not interrupt the discussion when there is a dominance event. An example scenario would be that one of the speakers has been speaking continuously reaching a threshold, without any interruption by the other participant. Whenever this turn length reaches a set threshold (estimated from the human recording data), the tutor interrupts the

- speaker by saying something like “But what do you think”, or “have you though that the moon does not have a magnetic field”, while turning the head towards the silent participant.
- Once all the cards are individually discussed, the Card Tracking module detects a final order of the card and generates a “cards ordering event”, moving the flow to the Order state. This state simply informs the speakers that they should start ordering the cards and moves to the Check Order state. When the system is in this state, silence periods are measured. If they reach the threshold, the system verifies if the order is correct. If the order is correct the system goes to the Confirm Order state and explicitly confirms if the final order has been reached and if both participants agree on the order the game ends. The Confirm Order state is the only state in the dialog that does not extend the behaviors of the Game state. The confirmation is made using speech recognition with a simple yes/no grammar. If the order is not correct, the system gives a transcribed hint that the human tutor used in this context. In this state, when the order timeout is reached, the system recaps the current order and goes to the Confirm Order state to explicitly confirm if that is the final order. This system will repeat these steps until the users agree on the order.
  - Similar to discussing the cards, the Active tutor performs the behaviors implemented in the Neutral tutor and above described, with two new features. Dominance is detected as described for the Discussion state and there is also a minimum ordering time threshold, that is, the tutor encourages the speakers to continue the ordering discussion if their discussion time is too short.

## 7 Discussion and Future Work

In this work, we presented a novel experimental setup, and corpus collection, and the design details of a complex multiparty dialogue system. One of the main criteria that dictated the design of the system is to keep the experimental setup as natural as possible, in order to allow users to employ natural behaviors common in human-human conversations. We also chose a task that would give the system a special role rather than trying to simulate a task-independent human-human multiparty dialogue. The choice we took in this project is to design a task that enforces certain restrictions on the interaction in a way that would give benefit to the technology employed rather than limit the interaction. The tutoring setup was set in a way to allow the tutor to give any type of information, or to choose to be passively monitoring the interaction, lowering the expectations of the users on the “apparent intelligence” of the tutor. The design of the task also employed the use of physical objects that could be tracked reliably using computer vision. This would produce solid pieces of information regarding the content of the dialogue, allowing the system to produce context-specific and information-rich content (such as giving hints about a specific card, whenever users are silent for a specific period of time).

The design of the dialogue system is also novel in several aspects. The system can handle two users at the same time, and take their visual and verbal activity into account. The choice of building a Conversational Dynamics component of such

interactions, we believe, is valid for a large set of face-to-face multiparty human-machine dialogues. Such setups target the development of human-like behaviors that will need to depend on large and long-term contexts and variables (such as dominance, involvement and engagement, etc.). Such modeling of high level conversational variables cannot be a simple extension of dialogue states.

From pilot tests with users, the system shows great potential. The ability of the system in knowing the addressee of a certain utterance produced by a user enhances the interaction significantly. We intend to carry a large user-study to evaluate the system thoroughly, in regards to conversational management strategies, using the Active and Neutral tutoring patterns and actions found in the corpus, which practically will tune the thresholds for different regulatory actions done by the robot.

The work established in this project can be regarded as a research platform to explore the effects of different conversational strategies on users. One can, for example, control certain parameters in Furhat's behavior, and tune them systematically to study large effects on the conversation, in an unprecedented way. The system for example, can attempt to bond with one user over the other by control of agreement, facial expressions, verbal and nonverbal feedback, and support with hints. The platform can also be used to study verbal and nonverbal alignment and entrainment in multiparty dialogue, where certain parameters (such as loudness, pitch, emotions, speech rate) can be controlled, and manipulated differently for each user.

The area of face-to-face multiparty dialogue is highly rich and unexplored, compared to its dyadic-dialogue counterparts. We think of this work as an attempt to design an experimental setup where different behaviors in face-to-face socially aware multiparty conversations can be studied.

**Acknowledgements.** This project was carried out at as part of the eNTERFACE'13 Multimodal Interfaces Workshop in Lisbon, Portugal. The project members are thankful to the organizers for providing the space and resources. Samer Al Moubayed was partly funded by the KTH Strategic Research Area - Multimodal Embodied Communication. Jekaterina Novikova is funded by the University of Bath and her participation was partly funded by the Society for the Study of Artificial Intelligence and Simulation of Behaviour. The authors would also like to thank the reviewers for their constructive comments, and the eNTERFACE participants for taking part in the data recordings.

## References

1. Cassell, J.: Embodied conversational agents. MIT Press, Cambridge (2009)
2. Rudnicky, A.: Multimodal dialogue systems. In: Minker, W., et al. (eds.) Spoken Multimodal Human-Computer Dialogue in Mobile Environments. Text, Speech and Language Technology, vol. 28, pp. 3–11. Springer (2005)
3. Clifford, N., Steuer, J., Tauber, E.: Computers are social actors. In: CHI 1994: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 72–78. ACM Press (1994)

4. Cohen, P.: The role of natural language in a multimodal interface. In: Proc. of User Interface Software Technology (UIST 1992) Conference, pp. 143–149. Academic Press, Monterey (1992)
5. Cohen, P., Oviatt, S.: The role of voice input for human-machine communication. *Proceedings of the National Academy of Sciences* 92(22), 9921–9927 (1995)
6. Castellano, G., Paiva, A., Kappas, A., Aylett, R., Hastie, H., Barendregt, W., Nabais, F., Bull, S.: Towards empathic virtual and robotic tutors. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 733–736. Springer, Heidelberg (2013)
7. Iacobelli, F., Cassell, J.: Ethnic Identity and Engagement in Embodied Conversational Agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 57–63. Springer, Heidelberg (2007)
8. Robins, B., Dautenhahn, K., te Boekhorst, R., Billard, A.: Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills? In: Universal Access in the Information Society, UAIS (2005)
9. Al Moubayed, S., Beskow, J., Skantze, G., Granström, B.: Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 114–130. Springer, Heidelberg (2012)
10. Skantze, G., Al Moubayed, S.: IrisTK: A statechart-based toolkit for multi-party face-to-face interaction. In: ICMI 2012, Santa Monica, CA (2012)
11. Oertel, C., Cummins, F., Edlund, J., Wagner, P., Campbell, N.: D64: A corpus of richly recorded conversational interaction. *Journal of Multimodal User Interfaces* (2012)
12. Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., House, D.: Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proc. of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), Valetta, Malta, pp. 2992–2995 (2010)
13. Al Moubayed, S., Edlund, J., Gustafson, J.: Analysis of gaze and speech patterns in three-party quiz game interaction. In: Interspeech 2013, Lyon, France (2013)
14. Paggio, P., Allwood, J., Ahlsen, E., Jokinen, K., Navarretta, C.: The NOMCO multimodal Nordic resource - goals and characteristics. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2010), Valetta, Malta (2010)
15. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation* 41(2), 181–190 (2007)
16. Digman, J.M.: Personality structure: Emergence of the five-factor model. *Annual Review of Psychology* 41, 417–440 (1990)
17. Bateman, T.S., Crant, J.M.: The proactive component of organizational behavior: A measure and correlates. *Journal of Organizational Behavior* 14(2), 103–118 (1993)
18. Langelaan, S., Bakker, A., Van Doornen, L., Schaufeli, W.: Burnout and work engagement: Do individual differences make a difference? *Personality and Individual Differences* 40(3), 521–532 (2006)
19. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) USAB 2008. LNCS, vol. 5298, pp. 63–76. Springer, Heidelberg (2008)
20. Cronbach, L.J.: Coefficient alpha and the internal consistency of tests. *Psychometrika* 16, 297–334 (1951)
21. Sacks, H.: A simplest systematics for the organization of turn-taking in conversation. *Language* 50, 696–735 (1974)

22. Duncan, S.: Some Signals and Rules for Taking Speaking Turns in Conversation. *Journal of Personality and Social Psychology* 23, 283–292 (1972)
23. Goodwin, C.: Restarts, pauses and the achievement of mutual gaze at turn-beginning. *Sociological Inquiry* 50(3–4), 272–302 (1980)
24. Bohus, D., Horvitz, E.: Facilitating multiparty dialog with gaze, gesture, and speech. In: *ICMI 2010, Beijing, China* (2010)
25. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. *Journal of Semantics* 9(1), 1–29 (1993)
26. Koutsombogera, M., Papageorgiou, H.: Linguistic and Non-verbal Cues for the Induction of Silent Feedback. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) *Second COST 2102. LNCS, vol. 5967*, pp. 327–336. Springer, Heidelberg (2010)
27. Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E., Koppensteiner, M.: The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behavior simulation. *Journal on Language Resources and Evaluation* 41(3–4), 255–272 (2007a)
28. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: A professional framework for multimodality research. In: *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 1556–1559 (2006)
29. 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 (2007b)
30. 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: *Seventh International Conference on Language Resources and Evaluation, LREC 2010* (2010)
31. Beskow, J.: Rule-based visual speech synthesis. In: *Proc of the Fourth European Conference on Speech Communication and Technology* (1995)
32. Al Moubayed, S., Edlund, J., Beskow, J.: Taming Mona Lisa: Communicating gaze faithfully in 2D and 3D facial projections. *ACM Transactions on Interactive Intelligent Systems* 1(2), 25 (2012)
33. Al Moubayed, S., Skantze, G.: Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. In: *AVSP 2011, Florence, Italy* (2011)
34. Al Moubayed, S., Skantze, G.: Perception of Gaze Direction for Situated Interaction. In: *4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, The 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA* (2012)
35. Al Moubayed, S., Skantze, G., Beskow, J.: Lip-reading Furhat: Audio Visual Intelligibility of a Back Projected Animated Face. In: *10th International Conference on Intelligent Virtual Agents (IVA 2012), Santa Cruz, CA, USA* (2012)
36. Skantze, G., Al Moubayed, S., Gustafson, J., Beskow, J., Granström, B.: Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue. In: *Proceedings of IVA-RCVA, Santa Cruz, CA* (2012)
37. Harel, D.: Statecharts: A visual formalism for complex systems. *Science of Computer Programming* 8(3), 231–274 (1987)
38. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: *Conference on Human Factors in Computing Systems*, pp. 858–859 (2002)



39. Ba, S.O., Odobez, J.-M.: Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 39(1), 16–33 (2009)
40. Johansson, M., Skantze, G., Gustafson, J.: Head Pose Patterns in Multiparty Human-Robot Team-Building Interactions. In: Herrmann, G., Pearson, M.J., Lenz, A., Bremner, P., Spiers, A., Leonards, U. (eds.) *ICSR 2013*. LNCS, vol. 8239, pp. 351–360. Springer, Heidelberg (2013)
41. Al Moubayed, S., Beskow, J., Granström, B.: Auditory-Visual Prominence: From Intelligibility to Behavior. *Journal on Multimodal User Interfaces* 3(4), 299–311 (2010)
42. Al Moubayed, S., Beskow, J.: Effects of Visual Prominence Cues on Speech Intelligibility. In: *Auditory-Visual Speech Processing, AVSP 2009*, Norwich, England (2009)
43. Streefkerk, B., Pols, L.C.W., ten Bosch, L.: Acoustical features as predictors for prominence in read aloud Dutch sentences used in anns. In: *Eurospeech*, Budapest, Hungary (1999)
44. Bevacqua, E., Pammi, S., Hyniewska, S.J., Schröder, M., Pelachaud, C.: Multimodal backchannels for embodied conversational agents. In: *The International Conference on Intelligent Virtual Agents*, Philadelphia, PA, USA (2010)
45. Zhang, J.Y., Toth, A.R., Collins-Thompson, K., Black, A.W.: Prominence prediction for super-sentential prosodic modeling based on a new database. In: *ISCA Speech Synthesis Workshop*, Pittsburgh, PA, USA (2004)
46. Al Moubayed, S., Chetouani, M., Baklouti, M., Dutoit, T., Mahdhaoui, A., Martin, J.-C., Ondas, S., Pelachaud, C., Urbain, J., Yilmaz, M.: Generating Robot/Agent Backchannels During a Storytelling Experiment. In: *Proceedings of (ICRA 2009) IEEE International Conference on Robotics and Automation*, Kobe, Japan (2009)
47. Terken, J.: Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America* 89, 1768–1776 (1991)
48. Wang, D., Narayanan, S.: An acoustic measure for word prominence in spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 690–701 (2007)
49. Kullback, S.: *Information Theory and Statistics*. John Wiley and Sons (1959)
50. Hotelling, H., Eisenhart, M., Hastay, W., Wallis, W.A.: *Multivariate quality control*. McGraw-Hill (1947)
51. Cheveigne, A.D., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111, 1917–1930 (2002)
52. Greenberg, S., Carvey, H., Hitchcock, L., Chang, S.: Temporal properties of spontaneous speech - Asyllable-centric perspective. *Journal of Phonetics* 31, 465–485 (2003)
53. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. *ACM Comput. Surv.* 38(4), Article 13 (2006)
54. Rienks, R., Heylen, D.: Dominance Detection in Meetings Using Easily Obtainable Features. In: Renals, S., Bengio, S. (eds.) *MLMI 2005*. LNCS, vol. 3869, pp. 76–86. Springer, Heidelberg (2006)



## Appendix B

**Table 4.** The annotation scheme employed for the manual analysis of the tutor conversational behavior

Annotation layers	Values
Speech_activity	Free text
Dialogue acts	Take, Accept, Grab, Offer, Complete, Yield, Hold, Back-channel
Turn management	Take, Accept, Grab, Offer, Complete, Yield, Hold, Back-channel
Feedback	Perception/Understanding (Give-Elicit) Accept (Give-Elicit), Non-accept (Give-Elicit)
Verbal_feedback	Free text
Face_general	Smile, Laugh, Scowl
Functions_Face	Feedback, Turn Management
Head_movement	Nod(s), Shake, Jerk, Tilt, Turn, Forward, Backward
Functions_Head_movement	Feedback, Turn Management
Gaze	Attention_Person_Right, Attention_Person_Left, Attention_Object, Glance
Functions_Gaze	Feedback, Turn Management
Eyes	Wide_open, Semi-closed, Wink, Blink
Functions_Eyes	Feedback, Turn Management
Eyebrows	Raise, Frown
Functions_Eyebrows	Feedback, Turn Management
Mouth	Open, Closed
Functions_Mouth	Feedback, Turn Management
Cards	Card id

# Touching Virtual Agents: Embodiment and Mind

Gijs Huisman<sup>1</sup>, Merijn Bruijnes<sup>1</sup>, Jan Kolkmeier<sup>1</sup>, Merel Jung<sup>1</sup>,  
Aduén Darriba Frederiks<sup>2</sup>, and Yves Rybarczyk<sup>3</sup>

<sup>1</sup> Human Media Interaction Group, University of Twente  
{gijs.huisman,m.bruijnes,m.m.jung}@utwente.nl,  
j.kolkmeier@student.utwente.nl

<sup>2</sup> Digital Life Centre, Amsterdam University of Applied Sciences  
{a.darriba.frederiks}@hva.nl

<sup>3</sup> New University of Lisbon  
yr@uninova.pt

**Abstract.** In this paper we outline the design and development of an embodied conversational agent setup that incorporates an augmented reality screen and tactile sleeve. With this setup the agent can visually and physically touch the user. We provide a literature overview of embodied conversational agents, as well as haptic technologies, and argue for the importance of adding touch to an embodied conversational agent. Finally, we provide guidelines for studies involving the touching virtual agent (TVA) setup.

**Keywords:** Embodied conversational agent, Touching virtual agent, Simulated social touch, Haptic feedback, Augmented reality.

## 1 Introduction

Embodied conversational agents (ECA) attempt to approximate human face-to-face communication through the use of, for instance, facial expressions, vocal expressions, and body postures. Different communication channels are used to give the virtual agent a more lifelike appearance. For example, an ECA can display realistic listening behavior by using head nods and vocal utterances (e.g. “uhuh”) while listening to a user telling a story [57]. Other examples include the use of facial expressions to express the agent’s emotional state [68], and the use of laughter by an agent to appear more human-like [93]. All of these signals affect the way users interact with the virtual agent [3]. However, one communication modality that is known to have strong effects on face-to-face communication between two human conversation partners, and that has been largely overlooked in ECAs, is touch [4]. Though social touch occurs less frequently between co-located individuals than other forms of social communication (e.g. head nods), it can have profound effects on the interaction [22][28]. For example, touch affects compliance to requests [34], can reduce stress [18], and can be used to communicate discrete emotions [42][43]. These effects

are strongly dependent on the context in which the communication takes place [12], such as the relation between conversation partners [11], the body location of the touch [46], the type of touch [24], and the communication partner's culture [67]. Effects can range from very positive affective responses, such as in the case of receiving a hug from a loved one, to very negative, for example when standing shoulder-to-shoulder in a busy train. Here we present a project in which we developed an ECA that has the capability to touch the user, while displaying a number of realistic touch behaviors (e.g. body posture, hand movement, and tactile sensation): a *touching virtual agent (TVA)*. We will refer to any system which uses some form of virtual representation of an agent, in combination with social touch capabilities by the agent, as a TVA.

The goal of the here presented project is twofold: first, by adding the tactile modality to an ECA we extend the communicative capabilities of the agent. We take another step towards ECAs that can display all the communicative subtleties of human-to-human communication, in an attempt to make communication with ECAs more lifelike. This has benefits for the use of ECAs in a range of scenarios such as training, therapy, and entertainment [13]. Second, a TVA (an ECA with social touch capabilities) would allow for the controlled study of tactile social behavior. Social touch by the agent could be studied in conjunction with other communication channels, such as facial expressions, and vocal utterances. The advantage of studying social touch using a TVA platform is that each communication channel can be minutely controlled, allowing for studies that disentangle the role of social touch in relation to other communication channels. This could provide valuable insights into the role of social touch as it occurs between co-located individuals, as well as social touch by virtual agents and social robots.

In the next section (Section 2) we outline related work on ECAs. We provide a brief overview of research into communication channels used in ECAs and describe different application areas. Furthermore, we outline research on the neurophysiology of touch, especially its relevance for social touch. Next, we provide an overview of work on the effects of social touch. Finally, we outline research on touch mediated by technology and early attempts of introducing touch to virtual characters. In Section 3 we describe our proposed TVA system. We outline design decision based on literature, and describe the features of the system, and potential application areas. Section 4 contains a detailed description of all the technical components of the system. In Section 5 we provide guidelines for experiments to be conducted with our TVA system. Section 6 conclude the paper, and provide suggestions for improvements of the system and future research directions.

## 2 Related Work

In this section we provide an overview of research on ECAs, dealing with specific communication modalities (e.g. facial and vocal expressions), effects of agent behaviors on communication with the agent, as well as examples of applications

of ECAs. In addition, we describe research on haptic perception, specifically dealing with social touch, effects of social touch on behavior, we give a brief overview of existing devices that mediate touch, and finally describe early work on TVAs.

### 3 Embodied Conversational Agents

Embodied (virtual) conversational agents should be able to display behaviors and skills similar to humans to be considered human-like. To accomplish this, they should meet the following six requirements as mentioned in the literature [25][47]. Social agents should be able to:

- 1 Express and perceive emotions;
- 2 Communicate with high-level dialogue;
- 3 Learn and recognize models of other agents;
- 4 Use natural cues (gaze, gestures, etc.);
- 5 Exhibit distinctive personality and character;
- 6 Learn and develop social competencies.

However, it is important to recognize what the agent will be used for. An agent that is to autonomously interact with humans (e.g. a receptionist) requires a bigger skill-set than an agent that is used in a controlled lab-experiment. In a lab setting, it is common that an experimenter controls the behavior of the agent through a “Wizard of Oz”-like setting and the agent only generates the behaviors selected by the experimenter.

In this chapter we focus on an agent for controlled experiments. However, to provide some more background on ECAs, we give a brief overview of the components required for an autonomously interacting agent.

#### 3.1 Complete the Loop

A human-like autonomous agent needs at least three general components: 1) some form of sensing. For example, sensing the environment and the (social) signals in this environment that are relevant. 2) mechanisms and models that can reason about the (social) environment and the role the agent has in this environment in order to come to goals the agent wishes to accomplish, and 3) a way to interact with the environment to attain its (social) goals. The actions by the agent influence the environment which leads to a new loop of perceptions, reasoning and actions by the agent.

**Perception.** An effective agent can perceive and comprehend its environment and interactants at a level that compares to that of real humans. This means that all human modalities need to be sensed by some sensor (e.g. camera for vision, microphone for sound, or a touch sensor for touch) and interpreted by some system.

Computer vision can provide object recognition [65] which can give the agent the ability to discuss the environment and make its conversational contributions grounded in the environment. Also, computer vision can give information about the social interaction itself, for example by recognizing human actions [80], detecting emotions from facial expressions [21], or recognizing the person that the system is engaged with [92].

Speech recognition can provide the agent with an understanding of the words or utterances made by the user, what the words (literally) mean, and what is meant with the words [75][82][91]. For example, if the agent requests something of the user and the user responds with the positive word “Sure”, it makes a world of difference whether the user responds with sarcastic or enthusiastic tone of voice.

Speech and visual recognition are two important modalities for virtual (conversational) agents to perceive but there are of course more. This chapter focuses on touch, but there is little work on agent perception of touch. Some preliminary work has been done by Nguyen et al. [74] where the hand of a user is detected in relation to a virtual agent’s body and the agent could give an (arbitrary) emotional response to a touch. Another example of a (robotic) agent that detects touch is [89], where a robotic pet can distinguish between different types of touch. The perception of touch by a TVA might be important for an autonomous agent, however the interpretation of the touch as well, would be paramount for meaningful tactile social interactions between an autonomous TVA and a user.

**Integration and Reasoning.** Hearing a string of words, seeing a set of objects, or feeling pressure on ones arm does not make a meaningful interaction. An agent needs to integrate the information from its different modalities and reason about the world around it to determine how to interact with the world.

The information that comes from different senses needs to be represented in some ‘world representation’, an information state [60]. Relating different ‘pieces’ of information is crucial to understand the world. The linking of information can have profound consequences. For example, the sound of a voice and the movements of lips are likely related, meaning that the words said by this voice are uttered by the person with the moving lips. The consequence is that this person is responsible for the content of the message and any appraisal of this message can be attributed to this person. So, if the system likes the message it could decide to like the person. Attributing words to a person and having an appraisal on what is said is not enough. Humans check if they understood the other and if they themselves are understood. The process of constructing a shared understanding is called “grounding”. Only when an interaction is grounded it becomes social as both parties are talking about the same thing, their utterances are influencing and interacting with each other in a meaningful manner [16].

In social interactions, humans have an understanding of the beliefs, intents and desires of the other. This *theory of mind* [81] can inform the appraisal of the other’s actions or lack of actions. For an agent it is important to have some

form of appraisal of the other’s actions to maintain a consistent and human-like interaction.

Theories on how to model social interactions in agents often come from psychology. For example, Leary’s interactional stance theory [61] describes the interaction between dominance and affiliation and how interactants influence each other on these dimensions. This theory has been used for modeling and motivating the social behavior of an artificial agent [9][10]. An agent that has emotions and a personality, that has an understanding of the world around it, and that can reason about this world is a more believable agent [3].

**Behavior Generation.** Once an agent ‘has made up its mind’ on what to say or do, it should have some way of expressing itself through its embodiment. A wide variety of agent embodiments exist ranging from very abstract (e.g. text-only dialog systems) to photo-realistic avatars with speech recognition, text-to-speech and real-time turn-taking [54].

The human communicative repertoire uses all the affordances of the human body. For example, we use our prehensile hands to gesture and touch, our eyes to (not) gaze at each other, our facial features to convey emotion, and modulate our voice to emphasize or clarify what we say [13]. An agent can but does not necessarily have to use the same repertoire when interacting. It might use text or abstract symbols to convey its message, it might use more realistic representations of these human communicative channels, or use a combination of abstract and realistic ways to communicate (for example text-to-speech and subtitles). For an agent to engage in social touch behavior however, it would need some way of ‘reaching out’ of the virtual world and entering a user’s personal space, in order to deliver a touch.

In this chapter we discuss adding the touch modality to the behavioral repertoire of an agent. The limited amount of work on TVAs shows that users are able to interpret the agent’s touch in terms of affect arousal and affect valence. Although other modalities are dominant, an agent that can touch can lead to a better relationship with the user [4].

### 3.2 Agent Frameworks

Several open-source frameworks are available to generate virtual human’s including their visualization, behavior and voice. It is becoming more and more easy to tailor these frameworks to specific needs. For example, Virtual Human Toolkit [40], ASAP [96], and using crowd sourcing [84]. In this chapter we discuss a touching virtual agent where the agent embodiment (upper body, head and voice) is generated using the ASAP framework [96]. The ASAP behavior realizer (a SAIBA compliant BML 1.0 realizer [87]) is capable of incremental behavior generation, meaning that new agent behavior can be integrated seamlessly with ongoing behavior.



### 3.3 Application Areas

ECAs are being used in many areas, ranging from training and coaching [10][40], to adherence to therapy [55] and social training [13], to receptionist tasks [58]. The very first ECA (a chatbot named ELIZA) had a therapeutic role that was intended to be a parody of an actual therapist [95]. It is clear that ECAs are broadly deployable. Research into new abilities of ECAs can result in ECAs being applied in new scenarios and for new purposes. In the current work we integrated the tactile modality into the communicative repertoire of an ECA, creating a TVA. In combination with other communication modalities, a TVA could be more successful in certain therapy settings. For example, a TVA could enhance expressions of empathy [4]. Furthermore, a TVA might be used in therapy with people suffering from Autism Spectrum Disorder who experience anxiety being in close proximity to, or being touched by others [26].

### 3.4 Touch: Tactual Perception

The sense of touch as it is employed in, for example, the exploration of objects with the hands, or when giving someone a hug, is comprised of different sensory sensations that, combined, are referred to as tactual perception [64]. Tactual perception can be functionally divided into cutaneous perception (also called tactile perception) and kinesthetic perception (also referred to as proprioception) [37][64].

Kinesthetic perception refers to the awareness of the relative position of one's limbs in time and space, as well as one's muscular effort [37][64] which mainly draws on information from mechanoreceptors in the skin and muscles, as well as from motor-commands executed in the brain [37]. Through kinesthetic perception, a person can, for example, get information about the size and weight of an object.

Cutaneous perception pertains to sensations that are a result of stimulation of different receptors in the skin. The human skin contains thermoreceptors that sense temperature, nociceptors that sense pain, and mechanoreceptors that sense deformations of the skin [37][51][64]. Relatively recent findings demonstrate that the hairy skin, such as that found on the forearm, contains a type of receptive afferent (i.e. nerve carrying information to the central nervous system), called C Tactile afferent (CT afferent), which is not found in the glabrous, or hairless skin, such as that found on the palms of the hand. These CT afferents respond particularly strongly to gentle, caress-like touches, resulting in a positively hedonic, pleasant perception [5][62][63]. Furthermore, CT afferents offer poor spatial and temporal resolution [5][76]. This has lead researchers to propose the social touch hypothesis, which states that the caressing touches to which CT afferents are sensitive, are particularly pertinent in affiliative interactions with other humans [5][62][63][76]. This may be especially the case between mothers and infants. Furthermore, CT afferents have not been found in genitalia, supporting the distinction between CT afferent touches and sexual functions [62]. Considering social touch, these findings are important in that they highlight

aspects of cutaneous perception that play a vital role in social interactions. CT afferents might serve as a “filter”, that operates in conjunction with other mechanoreceptors in order to determine if a certain touch has social relevance or not [70].

It has to be noted here that because cutaneous perception, in the form of mechanoreceptive sensations during, for example skin stretch, also plays a role in kinesthetic perception. Therefore cutaneous and kinesthetic perception can only be distinguished functionally, and not mechanically [64]. Indeed, situations where both cutaneous and kinesthetic perception play a central role are more common. This combination of cutaneous and kinesthetic perception is referred to as haptic perception [37][64].

Haptic perception is vitally important for the forming of one’s “body schema” which refers to a postural model that is continuously modified and updated whenever the body moves or adopts a new posture [27]. As the body schema is important in the organization of movement, it must integrate information of cutaneous (e.g. information from skin receptors) and kinesthetic (e.g. position of limbs) perception. In order to integrate information from the surface of the skin (i.e. cutaneous perception), with kinesthetic information, stimuli presented on the body’s surface must be transformed from locations on the body to locations in external space [27]. Moreover, multisensory integration (i.e. integration of information of other senses such as vision) can occur before the haptic sensation becomes conscious. This may either facilitate or impair one’s perception of the haptic stimulus, depending on the spatial localization [27]. A well known example of this is the ‘rubber hand’ illusion [6]. In this illusion, participants see a rubber hand in front of them that corresponds to their own left or right hand. When the participants’ hand is hidden, and the rubber hand as well as the hidden hand are stroked at the same time, and with the same velocity, most participants will experience the rubber hand as their own hand. When asked to point towards their hand participants who experienced the illusion will point more towards the rubber hand, an effect known as proprioceptive drift [6]. This finding lends strong support to the notion that the rubber hand is incorporated in the participants’ body schema. Important in the triggering of the illusion are congruent multimodal cues (i.e. tactile and visual) and prior internal body representations [66]. The strength of the illusion is strongly dependent on a first person perspective, synchronous stimulation, and an anatomically believable rubber hand [66].

These findings are highly relevant considering the multimodal nature of the TVA setup presented in this chapter. The augmented reality setup might be used to elicit visuotactile illusions similar to the rubber hand illusion. Here the question would be, to what extent participants experience their own arm, as viewed through the augmented reality setup, to be touched by the virtual character.

### 3.5 Touch: Social Touch

As was outlined in Section 3.4 the human skin contains specialized receptors that serve a selective function in distinguishing social touches from all other

kinds of touch. This highlights the importance of touch as a modality in social communication. Social touch has been found to play a role in the communication of support, appreciation, inclusion, sexual interest, affection, playfulness, and attention getting [52]. Moreover, a large body of experimental research in psychology has demonstrated a number of effects of social touch on social interactions.

One of the most thoroughly researched effects of social touch, pertains to compliance to requests [28][44]. The general premise is that when a person requests something from another person, while at the same time briefly touching the other person's arm, hand, or shoulder, the receiver of the request is more likely to comply to the request than when the request would be made without a touch. This effect has been demonstrated in numerous ecologically valid settings, and with recipients being either aware or unaware of the touch. One of the first studies on the role of touch in compliance found that when touched, uninformed participants were more likely (51%) to return a lost dime in a phone booth, than participants that were not touched (29%) [56]. Moreover, a different study found that a brief touch on the customer's forearm by the waitress, would increase the amount of money left as a tip by the customer in a bar [34]. Similar effects have been found in a restaurant, where a touch by the waiter or waitress to the customer's forearm increased the customer's compliance to menu item suggestions by the waiter or waitress [35]. Furthermore, a brief touch to the forearm increased the chances that people would spontaneously help a confederate to pick up dropped diskettes [33], and increased the chances that students would volunteer to write down an answer on the blackboard during class [31]. In addition, touch increases the duration that participants are willing to spend on a repetitive task, such as filling out bogus personality questionnaire items [78], or giving an opinion on difficult social issues [73]. In most of the cases described thus far, the touches were applied to the arm of the participant. Indeed, touches to the arm, and particularly the upper arm, may yield the strongest positive effect on compliance, though this effect may be mediated by the gender of the toucher and receiver of the touch [79]. Moreover, the type of request may be important in gaining compliance [98]. Touch may have little effect on requests that require psychologically more costly behaviors, such as signing up for blood donations [32].

In medical settings research has shown that a touch by a medical practitioner increased the chances that a patient would adhere to their medication [36]. When touches given to elderly clients by caregivers in an elderly home were combined with verbal encouragements, calorie and protein intake by the elderly increased [20]. Touch by a nurse prior to a patient's surgery, can decrease a patient's stress level, both measured through self-report and physiological measures [19][97]. Similar effects of touch on the reduction of stress in patients' in intensive care units have been observed [41]. Beneficial effects of touch in the alleviation of stress in health care settings, may be stronger for more intense types of touch. Message therapy has been found to be both successful in the release of pain, as well as the improvement of the mood of cancer patients [59].

For a more complete overview of the effects of message therapy the reader is referred to [23].

The effects of social touch, as outlined above, are numerous, and occur in diverse contexts. However, the exact nature of the underlying mechanisms responsible for these effects of social touch, are still unclear [28][44]. One explanation is that the touch signals to the recipient that the toucher likes, and trusts him/her. The perception of need, or the elicitation of positive feelings would in turn increase compliance [28][78]. However, this does not explain those situations in which the recipient is unaware of being touched. An alternative explanation is that the positive effects of social touch are related to the sensing of pleasant touch by CT afferent nerves and the encoding of social touch in the posterior insular cortex [28][69][70]. Still, contextual factors such as the relation between the recipient and the toucher, could play a vital role. Indeed, whereas touch in collaborative settings enhances helping behavior, touch in competitive settings can reduce helping behavior [12].

Finally, touch does not only communicate positive or negative affective states, but the nature of the touch itself can be used to communicate discrete emotions. Studies have shown that participants in Spain and the United States could decode certain discrete emotions communicated to them solely through touches to the forearm [43]. Participants could decoded anger, fear, disgust, love, gratitude, and sympathy at above chance levels, whereas happiness, surprise, sadness, embarrassment, envy, and pride were decoded at less than chance levels. Accuracy ranged from 48% to 83% which is comparable to the decoding accuracy of facial expressions. Specific touch behaviors were observed for communicating distinct emotions. For example, anger was mostly communicated by hitting, squeezing, and trembling, whereas love was mostly communicated by stroking, finger interlocking, and rubbing. Moreover, when the encoders of the touches were allowed to touch the decoders of the touches anywhere on their body, an additional two emotions, namely happiness and sadness, were decoded at above chance level [42]. However, in a recent reanalysis of the arm-only experiment [43] it was found that some of the recognition rates were dependent upon specific gender combinations in the encoder-decoder dyads [45]. This again indicates the importance of contextual factors in social touch.

### 3.6 Mediated Social Touch

Since two decades, researchers have attempted to communicate a sense of touch at a distance, using haptic feedback technology. This can be referred to as remote, or mediated social touch [37]. Mediated social touch can be defined as “the ability of one actor to touch another actor over a distance by means of tactile or kinesthetic feedback technology” [37, p. 153]. Generally speaking reasons to add touch to remote communication are, that the communication itself becomes richer through the addition of an extra modality [8][14], that touch is particularly well suited for intimate communication [77], and that touch can be used in situations where other modalities might be inappropriate or overly distracting [29].

With these reasons in mind, numerous devices have been constructed that allow for two or more people to touch each other at a distance. For example *inTouch* consists of two devices of three rollers each. When one person manipulates the rollers on their device the rollers on the second device move accordingly. The idea behind *inTouch* was to offer people separated by a distance a playful interaction possibility [8]. *ComTouch* was conceived as a way to add tactile communication to telephone conversations. A force sensitive resistor would communicate the finger pressure on a mobile phone to vibrotactile feedback on a second phone. This way the tactile channel could be used for turn-taking, mimicry, and adding emphasis during a phone conversation [14]. Similarly, the *Poke* system augments a mobile phone by adding an inflatable air bladder to the front and back of a mobile phone. During a phone call, force exerted on one bladder, results in inflation of the bladder on the other phone, which ‘pokes’ the cheek of the recipient. In a long-term study with couples in long-distance relationships, the *Poke* system was found to enhance emotional communication, and to add a sense of closeness to the partner [14]. Other systems that aim to provide a sense of closeness through mediated social touch are for example ‘hug over a distance’ [71], and ‘huggy pajama’ [90]. Both systems consist of a vest that allows a person to receive a hug at a distance from another person. A more general approach was taken with the design of the *TaSST* (Tactile Sleeve for Social Touch). The system consists of two sleeves each with a force sensitive input layer, and a vibration motor output layer. Touches to the input layer of one sleeve are felt as a vibration pattern on the second sleeve. The idea behind the *TaSST* was that it would allow users to communicate different types of touch at a distance [49].

Though many of the devices aimed at mediated social touch are promising, there has been a lack of empirical research into the effects of mediated social touch [37][48][94]. Still, studies have shown that mediated social touch, in the form of vibrotactile stimulation of the upper arm during a chat conversation, can have effects on compliance, that are similar to those in unmediated situations [38]. Using a similar ‘haptic messenger’ where participants could be touched on different body locations by either a same, or different gender conversation partner, it was found that mediated social touch was perceived as inappropriate when applied to certain body locations such as the stomach. Dyad composition in terms of gender did not have any effect [39]. Similar to studies on the tactile communication of emotions in co-located space [42][44], some studies have found that haptic feedback can be used to communicate diffuse affective states [88], or even discrete emotions using a force feedback joystick to simulate a handshake [2]. Moreover, the way people think about expressing emotions using mediated social touch, is relatively similar to what they would do to express those emotions in unmediated situations [50].

Mediated social touch is often used in conjunction with other communication channels, such as the visual and auditory channels. In a study, where mediated social touch in the form of a squeeze to the upper arm was used in conjunction with a verbal story, it was found that touches applied during emotional stories

enhanced feelings of closeness to the storyteller [94]. However, in a replication study investigating the effect of mediated social touch, at emotional and random moments during a verbal story on perceived social presence of the story teller, no effects of the touches were found [53]. Others found that touch applied in an immersive collaborative virtual environment did not enhance compliance to the request to sing [7]. Although this may be due to the psychological demanding nature of the request [32]. Conversely, a number of studies that used force feedback joysticks in collaborative virtual reality settings, found that the addition of haptic feedback of another person's actions enhances feelings of presence [15][30][85][86].

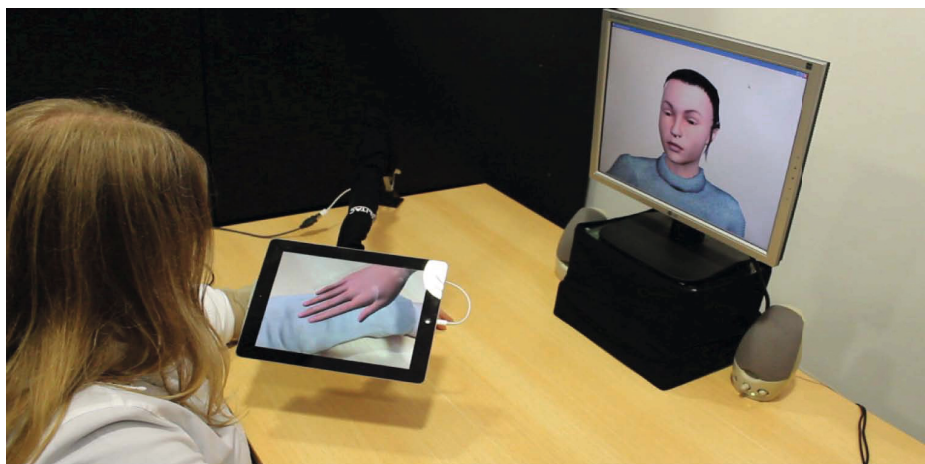
### 3.7 Touching Virtual Agents

The previous section outlined work on social touch mediated through haptic feedback technology. In most cases communication took place between dyads of participants, or participants were led to believe that they were communicating with another participant. However, the Media Equation theory suggests that humans interact with a perceived intelligent system as if the system is a human social actor [83]. Indeed, research into ECAs suggests that humans can engage in lifelike communication with virtual characters. Following this line of reasoning, touches applied by a TVA could be perceived by the user as social touches. Potential effects of co-located [28][44], and mediated social touch [37] might also apply here.

Some evidence for the notion that people indeed engage in realistic social touch behavior with virtual humans was found in a study into the way people touch digital objects [1]. Participants were tasked with brushing off 'dirt' from a virtual object using a force feedback joystick. Participants were presented with virtual humans, who would have dirt on their torso or face, and non-human objects such as geometric shapes. Results showed that participants touched virtual humans with less force than non-human objects, and that they touched the face of a virtual human with less force than the torso. Moreover, male virtual humans were touched with more force than female virtual humans [1].

In another study a TVA with a partially physical embodiment had the capability to squeeze a participant's hand through an air bladder system [4]. The TVA's face was displayed on a computer monitor situated on top of a mannequin body. The hand of the mannequin enclosed the participant's hand allowing the agent to squeeze it. In a series of experiments it was found that participants could perceive squeezes as expressions of affective arousal and valence by the agent. However, facial displays dominated the perception of affect. In an experiment where the agent expressed empathy for the participant, touches were found to enhance perceptions of the relation with the agent, but only for participants that felt comfortable being touched by others.

Social touch by ECAs, might be especially beneficial for interactions with physically embodied agents, such as social robots. In a study where participants observed videos of tactile interactions between a human and a social robot, it was found that social touch made the robot seem less machine-like and



**Fig. 1.** The components of the TVA setup. The monitor displays the agent's upper body, allowing the agent to gaze at, and talk to the user, as well as show body movements in accordance with the touch. The tablet computer displays the agent's hand touching the user. The user wears a vibrotactile sleeve to generate a tactile sensation in accordance with the agent's touch.

more dependable [17]. Other research, in which a social robot actually touched participants, found that social touch by a robot can enhance compliance to requests. After having their hand stroked by a robot, participants spent significantly more time on a repetitive task, and performed significantly more actions during the task [72].

## 4 Innovation

In the Section 2 we provided an overview of work on ECAs, and what is required to make them autonomously engage in human-like interactions. We detailed the concept of tactual perception, indicated that humans have specialized receptors for detecting social touches, and highlighted the importance of multimodal integration for haptic perception. In addition, we outlined some of the main effects social touch can have on interactions between co-located individuals. We drew parallels between social touch in co-located space, and social touch mediated through technology. Finally we described some early work on TVAs. In this section we built on the thus far discussed literature, to outline our TVA system.

The TVA setup consists of three components (see Figure 1) that are used in the social touch behavior of the agent. First, a visual representation of the upper body of the TVA. Second, using an augmented reality application, a visual representation of the hand of the TVA touching the user. And last, a tactile sensation that matches the location of the TVA's touch. To our knowledge, this is the first augmented reality TVA setup.

Where others have opted for a mannequin to represent the agent’s body and extremities [4], our setup combines an augmented reality display showing the agent’s hand with a monitor showing the agent’s body. This setup allows for animated movements of the agent’s body, arm, and hand to create a more realistic touching illusion. Moreover, by using a virtual representation of the agent rather than a physical representation, such as a social robot [72], we can incorporate more sophisticated facial expressions, as well as more easily manipulate the appearance of the agent. Also, the technical components required, are less cumbersome than those typically used in virtual reality [7], which requires a head mounted display and head tracking. As physical simulation of the agent’s touch we opted for a vibrotactile display, attached to the forearm. Vibrotactile stimulation is often used in mediated social touch devices [14][49], and has been found to have the potential to elicit effects similar to co-located social touch [38][39]. We chose to apply the touches to the forearm because it is an appropriate location for social touch to occur [79]. Moreover, we did not want the user to be focussed on the touch in itself, as is the case with, for example, a handshake [2] which is a very conscious and deliberate touch. A touch to the forearm may be less obtrusive than a handshake, potentially allowing for the study of the effects of touches of which the recipient is unaware [28]. In addition, touches to the forearm might be more appropriate as a form of affective expression [43]. Finally, the combination of visual and tactile feedback was considered to be important in creating the illusion of the agent touching the user. Research suggests that for, for example, the rubber hand illusion to occur congruent multimodal cues are important [66]. While the illusion created by the touching virtual agent system does not pertain to body ownership per se, congruent feedback from both the visible agent’s hand, and the tactile sensation might be important in creating a believable illusion of the agent’s touch.

In the next section we outline the technical architecture of the system, and provide technical details of all of the system’s components.

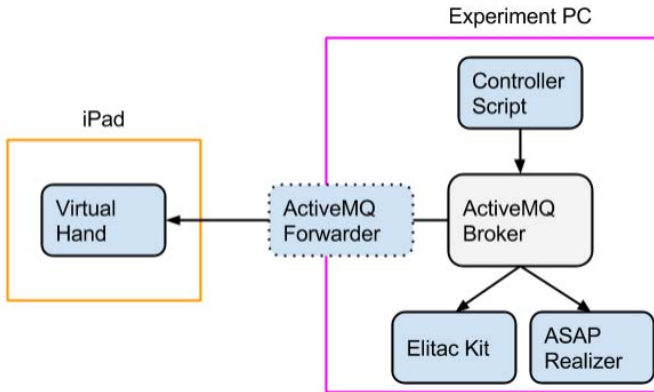
## 5 Technical Architecture

Building a TVA requires a virtual agent and a haptic device that can deliver the touching sensation. However, for the touch sensation to be believable, the integration of the the virtual world where the agent lives and the physical world where the user lives, is important. Specifically, the touch should consistently be attributed to the agent. This requires building consistent and believable agent behavior so that the touch is supported by the rest of the behavior (e.g. eye-gaze). Also, the setup of the entire agent system should allow for a believable presence of the agent. It should feel like the agent is actually there (albeit virtual).

### 5.1 A Virtual Touch

The agent’s touch illusion is created using the setup depicted in Figure 1. A user that is interacting with our TVA wears a vibrotactile sleeve on his/her left





**Fig. 2.** An overview of the system components

forearm. This arm is placed under a tablet, meaning the user sees his/her arm ‘through’ the tablet. The virtual agent is displayed on a computer screen placed behind and above the tablet. The tablet and the computer screen are aligned so that, for the user, the shoulder of the agent lines up with the user’s arm as seen on the tablet. When the agent touches the user, the agent first looks at the ‘to-be-touched’ area. Next, the shoulder and arm of the agent (on the computer screen) rotate and move forward towards the user’s arm. At the time that the agent’s virtual hand would reach the tablet’s view, a virtual hand and arm enter the tablet’s screen. When the virtual hand reaches the user’s hand, it stops and hovers for a moment before going in for the touch. At the moment of contact between the virtual hand and the user’s arm, the vibrotactile sleeve vibrates.

## 5.2 Components

The TVA system consists of four main components: a tablet computer, a vibrotactile sleeve, a virtual human, and a controller program. The components communicate over an ActiveMQ message broker (Figure 2).

**Controller.** The controller times and sends the commands to the different components for each agent action. Timing is important to create and maintain the illusion of a virtual presence that can touch the user. The controller is set up for easy manipulation of the timings of messages to different components. For example, if the tablet is moved further away from the computer screen the virtual hand should take slightly longer to reach the user, as the distance between the agent’s hand and body becomes longer. Thus the timing of the touch should be altered accordingly.



**Fig. 3.** The vibrotactile sleeve with 12 vibrationmotors

**Vibrotactile Sleeve.** To generate the tactile sensation of the agent's touch we used an Elitac Science Suit.<sup>1</sup> The Elitac Science Suit is a modular system consisting of several eccentric mass vibration motors that can be attached to elastic bands of different sizes using Velcro. The intensity of vibration of each vibration motor can be individually controlled, with four levels of vibration intensity available. We attached 12 vibration motors to an arm-sized elastic sleeve, with approximately half a centimeter spacing between each motor, creating a high resolution tactile display (Figure 3).

**Virtual Hand.** The tablet (iPad, 4th Generation, Apple Inc., Cupertino, California) runs a Unity project that can track the position of the user's arm and the position of the tablet relative to the table, and thus where the screen with virtual human is, using visual markers. Next, an augmented reality overlay is used to display a virtual hand and a shadow of this hand. The position of the tablet provides the information necessary to determine the starting point for the virtual hand. The position of the user's arm determines the target location for the virtual hand. Also, to give visual feedback of the distance of the virtual hand to the user's arm, the shadow of the virtual hand was overlaid on the user's arm.

<sup>1</sup> <http://elitac.org/products/sciencesuit.html>

**Virtual Human.** The virtual human is an instance of the ASAP-realizer [96]. The virtual human is controlled by sending BML-scripts that stipulate what it should do and when. The ASAP platform gives feedback on the progress and completion (or failure) of its actions. Such information might be used to time the actions of other components.

## 6 Research Applications

In this section we describe some of the benefits of our TVA setup. We focus specifically on the research domain. We discuss two future studies and provide the results of one pilot study, showing the feasibility of our setup as a research application.

### 6.1 Bug Demo

In the first technical demonstration of our system, the user can see a virtual bug walking on his/her arm. The tactile sleeve generates a tactile sensation of the bug walking on the arm. When the virtual agent detects the bug, she moves in to squash the bug on the user's arm. This slap is accompanied by an intense tactile stimulation.<sup>2</sup>

In this demo we investigated whether a “general” (i.e. discriminative) tactile sensation in the form of a bug walking on the arm can feel different from being touched by the agent. The participants in this demo reported that they felt the bug moving and that the tactile and visual stimuli created the illusion that a bug was walking on their arm. One participant stated he felt the urge to roll up the sleeve of his blouse because: “there was a big bug on my arm.” The virtual agent's slap was reported to be effective. Participants reported things like “she got it!”. This demo shows that the combination of visual and tactile stimuli can create the illusion of touch. The statements of the participants showed that they felt at least somewhat immersed in the scenario and as such could distinguish between discriminative touch and social touch.

### 6.2 Perception of Touch

The timing of the stimuli for the different modalities involved in detecting touch is a sensitive issue, as mentioned earlier. However, varying the timing of each stimulus can give us insights into elements that are important in ‘building’ certain touch behaviors. Incongruent visuotactile stimulation might ‘break’ the illusion of the TVA's touch. It would be of interest to investigate how much of a delay between the vibrotactile stimulation and the visual touch, would still constitute a touch be the TVA.

An extension of such a study, would be to present different types of touch (e.g. taps vs. strokes) to users. The modalities (visual and tactile) could be

---

<sup>2</sup> A video of this demo is available at: <http://vimeo.com/71094353>

either congruent or incongruent. For example, an incongruent stroke would be a stroking tactile stimulus combined with a visual tap. By varying the time of the behavior modalities and asking the participant to press a button as soon as they experience being touched, would give us insight into which modality is predominantly responsible for the perception of different touches. Also, it can show if the first notion of a touch (irrelevant whether this is visual or haptic) or if the complete package (user responds when the entire touch, both visual and tactile is perceived) is determinative.

### 6.3 Negotiation Task

As was outlined in Section 3.5, social touch can have numerous effects on social interactions. It would be interesting to investigate the role of social touch in a controlled manner with our TVA setup. De Melo et al. [68] used the iterated prisoners dilemma to investigate the effect of emotion on cooperation. In the iterated prisoners dilemma was described by [68] to their participants as: *“You and your partner are arrested by the police. The police have insufficient evidence for a conviction, and, having separated you both into different cells, visit each in turn to offer the same deal. If one testifies for the prosecution against the other and the other remains silent, the betrayer goes free and the silent accomplice receives the full 3-year sentence. If both remain silent, both prisoners are sentenced to only 3 months in jail for a minor charge. If each betrays the other, each receives a 1-year sentence”*. They showed that participants cooperate more with a virtual human that displays emotions and that they perceive the virtual human as more human-like. In conjunction with varying emotional displays, the virtual human could apply a touch. The question here would be: do touches by the virtual human, in combination with certain facial expressions, alter a participant’s decision in the prisoners dilemma?

### 6.4 Compliance Pilot Study

Social touch has been found to increase compliance to requests, both for co-located social touch [43] and mediated social touch [38]. Also, a previous study indicates that touch by a robot can increase the time and effort spend on a boring task [72]. In the current pilot study we wanted to see whether the TVA was able to increase compliance to a request when combining the request with a light touch on the arm. The TVA would make a request that the user perform a repetitive task, and combined this request with a touch to the arm of the user, or would make the request without a touch. For this study a task similar to [72] was constructed. The task consisted of the user using a computer mouse to drag a circle from the left side of the screen to a square on the right side of the screen. In the touch condition the TVA accompanied it’s request with a touch to the user’s forearm, and stated that the user could stop the task by pressing the “stop task button” that would be present on the screen. During the performance of the task every 20 seconds the TVA gave a verbal encouragement accompanied by a light touch to the user’s arm. In the no-touch

condition the same instructions and encouragements were given, but without the TVA touching the user. The task ended when the participant stopped the task by pressing the “stop task button”. To measure performance level we measured the number of circles which were successfully dragged to the square (counted as a single trial). We also measured the total amount of time spend on the task.

In a pilot test on 8 participants we found large individual differences. Most participants (5/8) quitted before the 5th trial, while the other participants continued up to 777 trials. Looking at the data the large differences were reflected in large standard deviations of completed trials in both the touch condition ( $M = 51.50, SD = 95.67$ ) and the no-touch condition ( $M = 203.50, SD = 382.56$ ). An independent samples t-test showed no significant differences between both conditions in completed trials ( $t(6) = .77, p = .47$ ) and time spend on the task ( $t(6) = .75, p = .48$ ).

Our results strongly contrast findings by [72], who found clear differences in performance when active touch (comparable to our touch condition) was applied compared to passive touch and no-touch. A potential explanation is that in [72], participant’s could more easily attribute the touch to the robot. In our TVA setup, this attribution might not always be as straightforward. Furthermore, cultural differences could explain the different findings. All participants in our pilot study were European, while the aforementioned study was conducted in Japan. During the experiment we also encountered some problems. First, the instructions were not clear to some of the participants. Despite the instruction given by the virtual agent it was unclear for some participants that *they* had to stop the task (2/8), whereas others tried the task one or two times, and then clicked the stop task button to see what would come next. Second, some participants had difficulty understanding what the virtual agent said (3/8). The pilot study would suggest that touch by a TVA might not affect compliance in the same manner as does co-located or mediated social touch. To further investigate this issue, a follow-up experiment would have to use a measure of trait compliance to tease out individual differences between participants. Furthermore, embedding the agent’s request in a more elaborate scenario might give the participants more time to get used to the agent’s voice, and touch, and the setup in general. Finally, a task such as filling out bogus personality items [78] might be considered more realistic by participants. It could be argued that participants would not simply try to fill out a few questions and see what comes next, as was the case with dragging the circles, but would take filling out the questionnaire more seriously, thus reducing initial differences between participants.

## 7 Conclusions

The TVA project set out to develop, from scratch, a setup in which a virtual character would be able to engage in communication with the user through facial expressions, speech, body movements, and, in this project most importantly, touch. The use of touch in social communication with an ECA has thus far been largely overlooked. An extensive body of work in psychology demonstrates

clear effects of social touch in co-located space on compliance, stress reduction, as well as communication of affect. Moreover, social touch mediated through haptic feedback technology has been found to have similar effects on compliance, and can be used to communicate emotions. Finally, some initial work on TVAs showed promising results.

Our approach differs from current efforts in creating touching virtual agents in that we combine a tactile sensation, with a visual representation of a hand touching the user in augmented reality, and other social communication from the agent (e.g. facial expressions and speech). With this setup we can introduce certain elements in the personal space of the user that could prompt the virtual agent to react and touch the user, as was for example the case in the bug demo. Moreover, the setup allows for the study of a number of effects of social touch by a TVA. We proposed experiments that could be conducted with the current setup. We conducted a pilot study in which we attempted to influence participant's compliance to a request made by the agent. Unfortunately, large individual differences in trait compliance negated any effects of the agent's touch.

Despite a current lack of empirical studies using the TVA setup, the setup does offer a good starting point for further exploration of touch by TVAs. We are confident that the combination of the virtual agent displaying other social signals (e.g. facial expressions and speech) with the application of touch using augmented reality and tactile sensations, is a viable approach to TVAs.

In this sense, eNTERFACE 2013, organized by the New University of Lisbon, Portugal, provided an excellent kick-start to the project. We came into eNTERFACE with a basic outline of the setup, required hardware, and ideas about the requirements for the setup. For this reason, much of the available project time was spent on constructing the setup, and working out all manners of technical difficulties. A preferable approach would have been to have already constructed some of the components, and then expand them during eNTERFACE, allowing time for conducting more studies with the setup. Nonetheless, the setup constructed allows for some interesting future studies, as described in Section 6, and offers a springboard for more elaborate TVA setups.

We suggest three specific improvements. First, the augmented reality area of the current setup is limited to the size of the tablet computer used (i.e. 9.7 inches). Expanding the visible augmented reality area would allow for a larger work-space in which the agent could interact with the user, and allowing the user more freedom of movement. Furthermore, a larger augmented reality area would enable us to display the agent's entire arm, making it easier for the user to attribute the touches to the agent. One approach would be to cover the area of the desk where the user's hands would be located, with a horizontally oriented opaque screen, with a wide-angle camera attached to the bottom of the screen. A projector could then project the image of the camera onto the opaque screen, expanding the augmented reality area. As a second improvement, using sensors, and possibly a fake limb in combination with an augmented reality overlay to simulate the agent's arm, we could allow the user to touch the agent. This would allow for bidirectional tactile communication. An interesting question in

this regard would be, if touch by the agent elicits similar tactile social behavior in the user. Third, the current setup makes use of a controller script that ties all of the separate system components together. A future refinement would be the construction of an ASAP engine to control all components of the system and replace the current control script. Specific BML commands could be defined, to make the system more robust and flexible.

**Acknowledgments.** This publication was supported by the Dutch national program COMMIT. Special thanks to Jan van Erp, Dirk Heylen, Ronald Poppe, and Dennis Reidsma, for their contributions to this work.

## References

1. Bailenson, J.N., Yee, N.: Virtual interpersonal touch: Haptic interaction and copresence in collaborative virtual environments. *Multimedia Tools and Applications* 37(1), 5–14 (2008)
2. Bailenson, J., Yee, N., Brave, S., Merget, D., Koslow, D.: Virtual interpersonal touch: Expressing and recognizing emotions through haptic devices. *Human-Computer Interaction* 22(3), 325–353 (2007)
3. Becker, C., Kopp, S., Wachsmuth, I.: Why Emotions Should be Integrated into Conversational Agents. *Engineering Approaches to Conversational Informatics*, pp. 49–68. John Wiley & Sons (2007)
4. Bickmore, T.W., Fernando, R., Ring, L., Schulman, D.: Empathic Touch by Relational Agents. *IEEE Transactions on Affective Computing* 1(1), 60–71 (2010)
5. Björnsdotter, M., Morrison, I., Olausson, H.: Feeling good: on the role of C fiber mediated touch in interoception. *Experimental Brain Research* 207(3-4), 149–155 (2010)
6. Botvinick, M., Cohen, J.: Rubber hands ‘feel’ touch that eyes see. *Nature* 391(6669), 756 (1998)
7. Bourdin, P., Sanahuja, J.M.T., Moya, C.C., Haggard, P., Slater, M.: Persuading people in a remote destination to sing by beaming there. In: *VRST 2013*, pp. 123–132. ACM (2013)
8. Brave, S., Dahley, A.: inTouch: A medium for haptic interpersonal communication. In: *Proceedings of CHI 1997*, pp. 363–364. ACM (1997)
9. Bruijnes, M.: Affective conversational models: Interpersonal stance in a police interview context. In: *Humaine Association Conference on Affective Computing and Intelligent Interaction, ASCII 2013*, pp. 624–629. IEEE Computer Society, USA (2013)
10. Bruijnes, M., Kolkmeier, J., op den Akker, R., Linssen, J., Theune, M., Heylen, D.: Keeping up stories: Design considerations for a police interview training game. In: *Social Believability in Games Workshop at ACE 2013* (2013)
11. Burgoon, J.K., Walther, J.B., Baesler, E.J.: Interpretations, evaluations, and consequences of interpersonal touch. *Human Communication Research* 19(2), 237–263 (1992)
12. Camps, J., Tuteleers, C., Stouten, J., Nelissen, J.: A situational touch: How touch affects people’s decision behavior. *Social Influence* 8(4), 237–250 (2013)
13. Cassell, J.: *Embodied conversational agents*. The MIT Press (2000)

14. Chang, A., O'Modhrain, S., Jacob, R., Gunther, E., Ishii, H.: Comtouch: Design of a vibrotactile communication device. In: Proceedings of DIS 2002, pp. 312–320. ACM (2002)
15. Chellali, A., Dumas, C., Milleville-Pennel, I.: Influences of haptic communication on a shared manual task. *Interacting with Computers* 23(4), 317–328 (2011)
16. Clark, H.H., Brennan, S.E.: Grounding in communication. *Perspectives on Socially Shared Cognition* 13, 127–149 (1991)
17. Cramer, H., Kemper, N., Amin, A., Wielinga, B., Evers, V.: Give me a hug: The effects of touch and autonomy on people's responses to embodied social agents. *Computer Animation and Virtual Worlds* 20(2-3), 437–445 (2009)
18. Ditzen, B., Neumann, I.D., Bodenmann, G., von Dawans, B., Turner, R.A., Ehlert, U., Heinrichs, M.: Effects of different kinds of couple interaction on cortisol and heart rate responses to stress in women. *Psychoneuroendocrinology* 32(5), 565–574 (2007)
19. Drescher, V.M., Gantt, W.H., Whitehead, W.E.: Heart rate response to touch. *Psychosomatic Medicine* 42(6), 559–565 (1980)
20. Eaton, M., Mitchell-bonair, I.L., Friedmann, E.: The effect of touch on nutritional intake of chronic organic brain syndrome patients. *Journal of Gerontology* 41(5), 611–616 (1986)
21. Fasel, B., Luetttin, J.: Automatic facial expression analysis: A survey. *Pattern Recognition* 36(1), 259–275 (2003)
22. Field, T.: Touch for socioemotional and physical well-being: A review. *Developmental Review* 30(4), 367–383 (2010)
23. Field, T., Diego, M., Hernandez-Reif, M.: Massage therapy research. *Developmental Review* 27(1), 75–89 (2007)
24. Floyd, K.: All touches are not created equal: Effects of form and duration on observers' interpretations of an embrace. *Journal of Nonverbal Behavior* 23(4), 283–299 (1999)
25. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. *Robotics and Autonomous Systems* 42(3), 143–166 (2003)
26. Foss-Feig, J.H., Heacock, J.L., Cascio, C.J.: Tactile responsiveness patterns and their association with core features in autism spectrum disorders. *Research in Autism Spectrum Disorders* 6(1), 337–344 (2012)
27. Gallace, A., Spence, C.: The cognitive and neural correlates of “tactile consciousness”: A multisensory perspective. *Consciousness and Cognition* 17(1), 370–407 (2008)
28. Gallace, A., Spence, C.: The science of interpersonal touch: An overview. *Neuroscience and Biobehavioral Reviews* 34(2), 246–259 (2010)
29. Gallace, A., Tan, H.Z., Spence, C.: The Body Surface as a Communication System: The State of the Art after 50 Years. *Presence: Teleoperators and Virtual Environments* 16(6), 655–676 (2007)
30. Giannopoulos, E., Eslava, V., Oyarzabal, M., Hierro, T., González, L., Ferre, M., Slater, M.: The effect of haptic feedback on basic social interaction within shared virtual environments. In: Ferre, M. (ed.) *EuroHaptics 2008*. LNCS, vol. 5024, pp. 301–307. Springer, Heidelberg (2008)
31. Guéguen, N.: Nonverbal encouragement of participation in a course: The effect of touching. *Social Psychology of Education* 7(1), 89–98 (2004)
32. Guéguen, N., Afifi, F., Brault, S., Charles-Sire, V., Leforestier, P.M., Morzedec, A., Piron, E.: Failure of Tactile Contact to Increase Request Compliance: The Case of Blood Donation Behavior. *Journal of Articles in Support of the Null Hypothesis* 8(1), 1–8 (2011)



33. Guéguen, N., Fischer-lokou, J.: Tactile contact and spontaneous help: An evaluation in a natural setting. *The Journal of Social Psychology* 143(6), 785–787 (2003)
34. Guéguen, N., Jacob, C.: The effect of touch on tipping: An evaluation in a french bar. *International Journal of Hospitality Management* 24(2), 295–299 (2005)
35. Guéguen, N., Jacob, C., Boulbry, G.: The effect of touch on compliance with a restaurant's employee suggestion. *International Journal of Hospitality Management* 26(4), 1019–1023 (2007)
36. Guéguen, N., Meineri, S., Charles-Sire, V.: Improving medication adherence by using practitioner nonverbal techniques: A field experiment on the effect of touch. *Journal of Behavioral Medicine* 33(6), 466–473 (2010)
37. Haans, A., IJsselsteijn, W.: Mediated social touch: A review of current research and future directions. *Virtual Reality* 9(2-3), 149–159 (2006)
38. Haans, A., IJsselsteijn, W.A.: The Virtual Midas Touch: Helping Behavior After a Mediated Social Touch. *IEEE Transactions on Haptics* 2(3), 136–140 (2009)
39. Haans, A., de Nood, C., IJsselsteijn, W.A.: Investigating response similarities between real and mediated social touch: A first test. In: *Proceedings of CHI 2007*, pp. 2405–2410. ACM (2007)
40. Hartholt, A., Traum, D., Marsella, S.C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., Gratch, J.: All together now. In: Aylett, R., Krenn, B., Pelachaud, C., Shimodaira, H. (eds.) *IVA 2013*. LNCS, vol. 8108, pp. 368–381. Springer, Heidelberg (2013)
41. Henricson, M., Ersson, A., Määttä, S., Segesten, K., Berglund, A.L.: The outcome of tactile touch on stress parameters in intensive care: A randomized controlled trial. *Complementary Therapies in Clinical Practice* 14(4), 244–254 (2008)
42. Hertenstein, M.J., Holmes, R., McCullough, M., Keltner, D.: The communication of emotion via touch. *Emotion* 9(4), 566–573 (2009)
43. Hertenstein, M.J., Keltner, D., App, B., Bulleit, B.A., Jaskolka, A.R.: Touch communicates distinct emotions. *Emotion* 6(3), 528–533 (2006)
44. Hertenstein, M.J., Verkamp, J.M., Kerestes, A.M., Holmes, R.M.: The communicative functions of touch in humans, nonhuman primates, and rats: A review and synthesis of the empirical research. *Genetic, Social, and General Psychology Monographs* 132(1), 5–94 (2006)
45. Hertenstein, M., Keltner, D.: Gender and the communication of emotion via touch. *Sex Roles* 64(1-2), 70–80 (2011)
46. Heslin, R., Nguyen, T., Nguyen, M.: Meaning of touch: The case of touch from a stranger or same sex person. *Journal of Nonverbal Behavior* 7(3), 147–157 (1983)
47. Heylen, D., op den Akker, R., ter Maat, M., Petta, P., Rank, S., Reidsma, D., Zwiers, J.: On the nature of engineering social artificial companions. *Applied Artificial Intelligence* 25(6), 549–574 (2011)
48. Huisman, G.: A touch of affect: Mediated social touch and affect. In: *ICMI 2012*, pp. 317–320. ACM (2012)
49. Huisman, G., Darriba Frederiks, A., Van Dijk, E., Heylen, D., Kröse, B.: The TaSST: Tactile Sleeve for Social Touch. In: *Proceedings of WHC 2013*, pp. 211–216. IEEE (2013)
50. Huisman, G., Darriba Frederiks, A.: Towards tactile expressions of emotion through mediated touch. In: *CHI 2013 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2013*, pp. 1575–1580. ACM (2013)
51. Johnson, K.O.: The roles and functions of cutaneous mechanoreceptors. *Current Opinion in Neurobiology* 11(4), 455–461 (2001)

52. Jones, S.E., Yarbrough, A.E.: A naturalistic study of the meanings of touch. *Communication Monographs* 52(1), 19–56 (1985)
53. Jung, M., Boensma, R., Huisman, G., Van Dijk, E.: Touched by the storyteller: the influence of remote touch in the context of storytelling. In: *Proceedings of ACII 2013*, pp. 792–797. ACM (2013)
54. Khan, R., De Angeli, A.: Mapping the demographics of virtual humans. In: *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... But Not as We Know It*, vol. 2, pp. 149–152. British Computer Society (2007)
55. Klaassen, R., op den Akker, H., Lavrysen, T., van Wissen, S.: User preferences for multi-device context-aware feedback in a digital coaching system. *Journal on Multimodal User Interfaces*, 21 (2013)
56. Kleinke, C.L.: Compliance to requests made by gazing and touching experimenters in field settings. *Journal of Experimental Social Psychology* 13(3), 218–223 (1977)
57. de Kok, I., Heylen, D.: Integrating backchannel prediction models into embodied conversational agents. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) *IVA 2012*. LNCS, vol. 7502, pp. 268–274. Springer, Heidelberg (2012)
58. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A conversational agent as museum guide—design and evaluation of a real-world application. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) *IVA 2005*. LNCS (LNAI), vol. 3661, pp. 329–343. Springer, Heidelberg (2005)
59. Kutner, J.S., Smith, M.C., Corbin, L., Hemphill, L., Benton, K., Mellis, B.K., Beaty, B., Felton, S., Yamashita, T.E., Bryant, L.L., Fairclough, D.L.: Massage therapy versus simple touch to improve pain and mood in patients with advanced cancer randomized trial. *Annals of Internal Medicine* 149(6), 369–379 (2008)
60. Larsson, S., Traum, D.R.: Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering* 6(3&4), 323–340 (2000)
61. Leary, T.: *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation*. Ronald Press, New York (1957)
62. Liu, Q., Vrontou, S., Rice, F.L., Zylka, M.J., Dong, X., Anderson, D.J.: Molecular genetic visualization of a rare subset of unmyelinated sensory neurons that may detect gentle touch. *Nature Neuroscience* 10(8), 946–948 (2007)
63. Löken, L.S., Wessberg, J., Morrison, I., McGlone, F., Olausson, H.: Coding of pleasant touch by unmyelinated afferents in humans. *Nature Neuroscience* 12(5), 547–548 (2009)
64. Loomis, J.M., Lederman, S.J.: Tactual perception. In: Boff, K., Kaufman, L., Thomas, J. (eds.) *Handbook of Perception and Human Performance: Cognitive Processes and Performances*, vol. 2, ch. 31, pp. 1–41. Wiley, NY (1986)
65. Lowe, D.G.: Object recognition from local scale-invariant features. In: *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157. IEEE (1999)
66. Maselli, A., Slater, M.: The building blocks of the full body ownership illusion. *Frontiers in Human Neuroscience* 7(83) (2013)
67. McDaniel, E., Andersen, P.: International patterns of interpersonal tactile communication: A field study. *Journal of Nonverbal Behavior* 22(1), 59–75 (1998)
68. de Melo, C.M., Zheng, L., Gratch, J.: Expression of moral emotions in cooperating agents. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) *IVA 2009*. LNCS, vol. 5773, pp. 301–307. Springer, Heidelberg (2009)

69. Morrison, I., Björnsdotter, M., Olausson, H.: Vicarious responses to social touch in posterior insular cortex are tuned to pleasant caressing speeds. *The Journal of Neuroscience* 31(26), 9554–9562 (2011)
70. Morrison, I., Löken, L., Olausson, H.: The skin as a social organ. *Experimental Brain Research* 204, 305–314 (2010)
71. Mueller, F.F., Vetere, F., Gibbs, M.R., Kjeldskov, J., Pedell, S., Howard, S.: Hug over a distance. In: *Proceedings of CHI 2005*, pp. 1673–1676. ACM (2005)
72. Nakagawa, K., Shiomi, M., Shinozawa, K., Matsumura, R., Ishiguro, H., Hagita, N.: Effect of robot's active touch on people's motivation. In: *HRI 2011*, pp. 465–472. ACM (2011)
73. Nannberg, J.C., Hansen, C.H.: Post-compliance touch: An incentive for task performance. *The Journal of Social Psychology* 134(3), 301–307 (1994)
74. Nguyen, N., Wachsmuth, I., Kopp, S.: Touch perception and emotional appraisal for a virtual agent. In: *Proceedings Workshop Emotion and Computing-Current Research and Future Impact, KI*, pp. 17–22 (2007)
75. Nygaard, L.C., Queen, J.S.: Communicating emotion: Linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance* 34(4), 1017 (2008)
76. Olausson, H.K., Wessberg, J., Morrison, I., McGlone, F., Vallbo, A.: The neurophysiology of unmyelinated tactile afferents. *Neuroscience and Biobehavioral Reviews* 34(2), 185–191 (2010)
77. Park, Y.W., Nam, T.J.: Poke: A new way of sharing emotional touches during phone conversations. In: *CHI 2013 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2013*, pp. 2859–2860. ACM (2013)
78. Patterson, M.L., Powell, J.L., Lenihan, M.G.: Touch, compliance, and interpersonal affect. *Journal of Nonverbal Behavior* 10(1), 41–50 (1986)
79. Paulsell, S., Goldman, M.: The effect of touching different body areas on prosocial behavior. *The Journal of Social Psychology* 122(2), 269–273 (1984)
80. Poppe, R.: A survey on vision-based human action recognition. *Image and Vision Computing* 28(6), 976–990 (2010)
81. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1(04), 515–526 (1978)
82. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2), 257–286 (1989)
83. Reeves, B., Nass, C.: *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places*. CSLI Publications (2002)
84. Rossen, B., Lok, B.: A crowdsourcing method to develop virtual human conversational agents. *International Journal of Human-Computer Studies* 70(4), 301–319 (2012)
85. Sallnäs, E.-L.: Haptic feedback increases perceived social presence. In: Kappers, A.M.L., van Erp, J.B.F., Bergmann Tiest, W.M., van der Helm, F.C.T. (eds.) *EuroHaptics 2010, Part II. LNCS*, vol. 6192, pp. 178–185. Springer, Heidelberg (2010)
86. Sallnäs, E.L., Rasmus-Gröhn, K., Sjöström, C.: Supporting presence in collaborative environments by haptic force feedback. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7(4), 461–476 (2000)
87. Schlangen, D., Skantze, G.: A general, abstract model of incremental dialogue processing. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 710–718. Association for Computational Linguistics (2009)

88. Smith, J., MacLean, K.: Communicating emotion through a haptic link: Design space and methodology. *International Journal of Human-Computer Studies* 65(4), 376–387 (2007)
89. Stiehl, W.D., Breazeal, C.: Affective touch for robotic companions. In: Tao, J., Tan, T., Picard, R.W. (eds.) *ACHI 2005*. LNCS, vol. 3784, pp. 747–754. Springer, Heidelberg (2005)
90. Teh, J.K.S., Cheok, A.D., Peiris, R.L., Choi, Y., Thuong, V., Lai, S.: Huggy pajama: A mobile parent and child hugging communication system. In: *IDC 2008*, pp. 250–257. ACM (2008)
91. Truong, K.P., Heylen, D.: Disambiguating the functions of conversational sounds with prosody: The case of ‘yeah’ (2010)
92. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: *Proceedings CVPR 1991*, pp. 586–591. IEEE (1991)
93. Urbain, J., Niewiadomski, R., Hofmann, J., Bantegnie, E., Baur, T., Berthouze, N., Cakmak, H., Cruz, R.T., Dupont, S., Geist, M., et al.: Laugh machine. *Proceedings eNTERFACE 12*, 13–34 (2013)
94. Wang, R., Quek, F., Tatar, D., Teh, K.S., Cheok, A.: Keep in touch: Channel, expectation and experience. In: *Proceedings of CHI 2012*, pp. 139–148. ACM (2012)
95. Weizenbaum, J., McCarthy, J.: Computer power and human reason: From judgment to calculation. *Physics Today* 30, 68 (1977)
96. van Welbergen, H., Reidsma, D., Kopp, S.: An incremental multimodal realizer for behavior co-articulation and coordination. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) *IVA 2012*. LNCS, vol. 7502, pp. 175–188. Springer, Heidelberg (2012)
97. Whitcher, S.J., Fisher, J.D.: Multidimensional reaction to therapeutic touch in a hospital setting. *Journal of Personality and Social Psychology* 37(1), 87–96 (1979)
98. Willis, F.N., Hamm, H.K.: The use of interpersonal touch in securing compliance. *Journal of Nonverbal Behavior* 5(1), 49–55 (1980)

**Part II**  
**Key Applications**

# Kinect-Sign: Teaching Sign Language to “Listeners” through a Game

João Gameiro, Tiago Cardoso, and Yves Rybarczyk

Universidade Nova de Lisboa,  
Faculdade de Ciências e Tecnologia  
Monte da Caparica, Portugal

j.gameiro@campus.fct.unl.pt, {tomfc, yr}@uninova.pt

**Abstract.** Sign language is the hearing impaired form of communicating with other people, including listeners. Most cases, impaired people have learned sign language from childhood. The problem arises when a listener comes in contact with an impaired person. For instances, if a couple has a child which is impaired, the parents find a challenge to learn the sign language. In this article, a new playful approach to assist the listeners to learn sign language is proposed. This proposal is a serious game composed of two modes: *School-mode* and *Competition-mode*. The first offers a virtual school where the user learns to sign letters and the second offers an environment towards applying the learned letters. Behind the scenes, the proposal contains a sign language recognition system, based on three modules: 1 – the standardization of the Kinect depth camera data; 2 – a gesture library relying on the standardized data; and 3 – the real-time recognition of gestures. A prototype was developed – Kinect-Sign – and tested in a Portuguese Sign-Language school and on eNTERFACE’13 resulting in a joyful acceptance of the approach.

**Keywords:** Kinect Sensor, Sign Language, Serious Game, Gesture Recognition.

## 1 Introduction

Communication is the most vital mean for understanding and being understood. Nevertheless, hearing impaired persons are faced with barriers that prevents communication. According to the World Health Organization,  $360 * 10^6$  persons around the globe suffer from disabling hearing loss. From those,  $328 * 10^6$  are adults, where the rest are children. According to the same organization, approximately one-third of the population over 65 years of age suffer from disabling hearing loss [1].

Most of the hearing impaired knows sign language, being developed by this group through time. But a problem emerge when non-impaired persons try to communicate with an impaired person, for instances, in the case of family, friends and colleagues, of the impaired. This happens since there is almost no learning mechanism designed for listeners.

As summarized in [2], the research community has placed a great effort in developing a sign language recognition (SLR) system, but, as the authors say, the problem is far from being solved. Most implementations rely on image processing, and, despite

the current advances on matching algorithms, still exists a need for great processing power in order to support real-time recognition. Nevertheless, the research community has been focused on showing that it is possible to make SLR, but limiting their studies in terms of lexicon or relying on the usage of special gloves, as presented in [3].

In what concerns the commercial aspect of teaching this form of communication, there are not many forms one can learn sign language. Nevertheless, specialized schools and videos exist, but are still limited.

The Kinect device, introduced by Microsoft, was intended to revolutionize the gaming industry [4], with the removal of the joystick/controller from the game. Alongside this revolution, the Kinect sensor also started appearing in other research areas. One of this is the SLR area, where the research community started changing from traditional camera-based approaches to Kinect-based approaches [5], [6].

The main goal of this article is to present an extension to the Kinect SDK, which provides gesture handling support, and proposes the creation of a serious game for listeners to learn sign language.

## 2 State of the Art

Teaching sign language through a serious game is a project that involves distinct areas. The first area of interest, a cornerstone of the proposal, is sign language, where there is a need to evaluate features, algorithms for recognition and existing games. The second area is the Kinect device, from which it is required to understand how it works and supported games.

### 2.1 Sign Language

Sign language shares, as much, similar aspects with spoken languages, as differences. For example, while spoken languages uses sound patterns to convey meaning, sign language is based on the usage of hand patterns and body language. Despite this difference both languages are considered natural languages and some of the sign languages have even achieved legal recognition [7], such as the Portuguese Sign Language (PSL). The Portuguese sign language alphabet is shown in Figure 1.



**Fig. 1.** Manual alphabet of the Portuguese Sign Language [6]

**Sign Language Games.** As stated before, according to the World Health Organization (2013), the hearing impaired community is relatively small and, as a result, there is not much offer in what concerns games using sign language. Despite this, it is possible to find some games, in various types of platforms.

Three examples, which are available on the market are:

- *Sign Language Bingo*, a board game with 201 basic sign vocabulary words [8].
- *Sign-O*, a computer game, also based on bingo, with 12 categories and where each board contains 25 words [9]. The Figure 2b represents the cover for the CD.
- *Sign the Alphabet*, an internet game that is played by identifying the letters and numbers that are shown [10].



**Fig. 2.** Image with three different games available for teaching Sign Language

“Serious games have become a key segment in the games market as well as in academic research” [11]. Therefore, the games presented in Figure 2, can be useful when learning sign language.

Nevertheless, there is still room for improvements, for instance, on one hand, most games are based on Bingo and, on the other hand, there is an inability to correctly evaluate the signs made. This makes the computer games not truly interactive and, therefore, makes some people say that a truly sign language interactive game would be a good game to buy [12].

**Sign Language Algorithms.** SLR has been of great interest on the research community for the past 20 years, for example the work of Starner in 1995 [13]. This is due to the characteristics of sign language, but, despite this interest, the improvements made on SLR were still very limited, generally recognizing up to ten signs and where not able to reach the hearing impaired community.

Until recently, this field of study have been losing importance in the research community. But with the appearance of the Kinect sensor a change occurred in this field. This is justified by the development of a depth camera capable of detecting distances in relation to the sensor. Other contributing fields are the machine learning and the computer vision fields, which allowed the detection of the human skeleton, with 20 joints, through the depth camera.

Until recent years, most of the SLR systems were based on Hidden Markov Models, a statistical model in which the system to be modelled is assumed as a Markov



process with hidden states, as shown in [14]. The work of Starner is based on this models, where he uses a single camera for data acquisition and the signing is tracked with the assistance of solid colour gloves. The prototype developed achieved a 97 percent accuracy while sing a forty word lexicon [13].

Recently, also due to the development on the machine learning and computer vision fields, Correia proposed two different algorithms for SLR: 1 – a K-Nearest Neighbour (KNN); and 2 – a Support Vector Machine (SVM). Just like in Starner' work, Correia was able to achieve good accuracy results, 96.58 percent for the KNN and 92.06 percent for the SVM, but he's work was limited to just four letters [15].

One of the most recent studies in SLR was developed in cooperation between Key Lab of Intelligent Information Processing, from China, with Microsoft, where the implemented prototype is based on a three dimensional trajectory matching, assisted by the Kinect sensor. This trajectory matching is made by acquiring the trajectory of the user hand, normalizing this trajectory, by linear resampling, and comparing with the trajectories gallery they possess. The study achieved a 96.32 percent accuracy using a gallery with 239 [5].

Based on this three systems it's possible to determine that high accuracy rates have been achieved in those studies, but the great majority of the studies are very limited in terms of amount of elements recognizable and there is still the matter of transition between academic study and market product.

## 2.2 Kinect Sensor

Code named *Project Natal*, the Kinect sensor was released on November 4, 2010, born from the collaboration between Microsoft and, the Israeli company, Prime Sense [4]. Their intent, for Kinect sensor, was to revolutionize the world of games, opening what they called new horizons and domains. To achieve it they alienated the standard controller and joystick. Instead, through the Kinect sensor, the natural movements of the body are used to control the form in which the game is played.

The introduction of the third dimension on the computer vision field, through a depth camera, is one of the key elements that assist in the removal of the game controller/joystick. This third dimension, or distances from the sensor, rely on the eco of infrared lights, or in other words, the measurement of distortions between distinct infrared beams, emitted by the sensor. Moreover, at a firmware level, other key element, the Kinect sensor is capable of providing skeleton and facial tracking. One last element, which allowed the first two to work properly, is the development of machine learning techniques.



**Fig. 3.** Composition of the Kinect sensor: (1) - Infrared optics; (2) - RGB camera; (3) - Motorized tilt; (4) - Multi-array microphone [4]

The Kinect sensor is mainly composed, as shown in Fig. 3, of the following four components:

- 1) The *depth camera*, or infrared optics, responsible for understanding the 3D environment in front of the Kinect sensor.
- 2) The *RGB camera*, which besides showing the user on screen it’s also responsible to make the facial recognition of the user.
- 3) A *motorized tilt*, mechanical gear that let the sensor follow the user.
- 4) A *multi-array microphone*, composed of four microphones embedded inside the Kinect sensor, is capable of isolate the user voice from the rest of the noise. It’s also capable of pinpointing the user location in relation to the Kinect sensor.

Besides this four visible components, the Kinect sensor possesses its own processor and firmware. As stated before, this is one of the key elements for the working of the Kinect sensor, since they are responsible for identifying the information acquired by the infrared optics. In other words, they have the ability to “build the world” visible by the sensor and, from that “world”, it’s capable to extrapolate 20 joints in the human body [4].

**Kinect Games.** Kinect sensor was initially design to support games, and, since it’s been released, many games have been created, for example *Kinect Sports* and *Star wars*, using this innovative device [16].

Another game, that appeared using the Kinect sensor, is *Minute To Win It*. In this game the user has 60 seconds to perform a simple task, and for every task he completes, he will earn a certain amount of money which can lead up to \$1 million [17], shown in Fig. 4.



Fig. 4. Screenshot of the game "*Minute To Win It*" [17]

### 3 Innovation - Multimodal

This project is developed around the Kinect sensor, so through the usage of this sensor a great level in innovation on the multimodal interfaces is already achieved. The reason for that is the lack of games and applications that uses this type of device. Also recognizing sign language through the Kinect sensor can be viewed as innovative, since we are recognizing the entire alphabet.

In the project, the main innovation is the creation of a serious game used to teach and, also, to be played with sign language. This is a brand new area, where there is nothing to compare, as it was visible in chapter 2.1. So it's possible to affirm that this types of games will develop into a brand new dimension.

This project is composed, mainly, by two components, the SLR system and the sign language game. The first can be expressed as the core for the entire project, since it's one of the most important systems that are embedded in the game. The SLR system will be studied in great detail so that it might be achieved the best results possible, before the usage of this system in the game. The second component, the sign language game, offers the players lessons where one can learn letters in sign language and offers some games where those lessons can be applied.

This project has a wide range of applications, starting from a classroom, where the teacher is assisted by the game to teach and to correctly evaluate the sign the student do. Or, in a more familiar environment, this game can be played between family members, assisting in the creation and development of bonds.

## 4 Proposal

The main goal on this section is to propose a serious game – Kinect-Sign – to teach sign language to listeners and to enrol the users in games on the sign language taught. For that reason, the game is designed with two different modes: *School-mode* and *Competition-mode*. Prior to the development of the serious game itself, there was the need to create a SLR algorithm that would work in the simplest and fastest way. Therefore, despite the already existing algorithms, it is proposed a very simple recognition algorithm to be used in the game.

### 4.1 Sign Language Recognition

The proposed SLR is divided in three components: 1 – the data standardization, that consists on the standardization of the acquired data through the Kinect sensor; 2 – the data storage, responsible for the creation of a sign language gestures library; and 3 – the data recognition, where the acquired depth data, from the Kinect sensor, is matched with the existing data in the sign language gesture library.

**Data Standardization.** The first step, prior of storing or matching any data, is the standardization of the acquired data. In other words, set the acquired data into a specific format for which standardizes the recognition process. This selected format is a 144×144 grayscale bitmap. In order to acquire this format the data acquired from the Kinect sensor goes through the following 6 steps:

1. Acquire the raw depth data from the Kinect sensor.
2. Obtain, from the skeleton frame (provided by the Kinect SDK), the right hand point. This point is the centre of the region of interest (ROI).
3. From the right hand point, determine the size of the ROI, according to the distance of the hand to the Kinect sensor.

4. After acquiring the size of the ROI, determine the corners for this ROI and then the depth image is split to get just the ROI.
5. Convert the depths in the ROI into a grayscale.
6. Concluding with scaling the grayscale bitmap to 144×144, through a nearest neighbour interpolation.

**Data Storage.** Data storage is the component responsible for storing the sign language library. This library is composed of the grayscale bitmaps, stored with two different purposes: 1 – a source data to use in the recognition process; and 2 –a test data, used in the validation of the SLR algorithm. Also, this data can be used for future developments.

According to the proposed purposes the data is divided into two categories. First, there is single bitmaps, or images, used both on the source data and test data. For the source data it will be stored 300 bitmaps per letter of the alphabet, while for the test data it is only stored 6 distinct bitmaps. Second, there is videos, composed of 300 bitmaps, stored in the test data, as to simulate real-time recognition. So, there will be three videos per letter of the alphabet, for recognition purposes.

**Data Recognition.** The last component is responsible for the matching between the gesture library and the Kinect acquired data. So, the proposed is an algorithm responsible for SLR. This occurs by acquiring a new image, with the Kinect sensor, and comparing it with all the images in the library. In order to implement this algorithm, the centrepiece for making the recognition is used the distance between depths, or in the prototype case, since the depths are converted into a grey scale, the distance between the greys.

Even though this evaluation of distances might be a good idea, looking to the standard format of the images and how the images are, Fig. 5, it is possible to get to two conclusions: 1 – the recognition achieves great accuracies, due to the great amount of black pixels in the image acquired; and 2 – being the images set to 144×144 pixels, there will be 20.736 pixels to match, which can be too expensive for most computers available.

Bearing this in mind, it is devised five different conditions to make the recognition and to assist in determining the most accurate algorithm possible, also developing a lighter algorithm. Those five conditions are shown in Table 1, and can be viewed as algorithms to ignore the black pixels of the images.

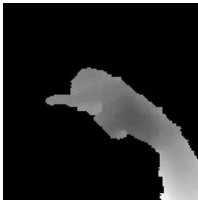
**Table 1.** Different conditions to ignore black pixels when using the recognition system

<i>Condition</i>	<b>Kinect</b>	<b>Library</b>
<i>None</i>	No	No
<i>Kinect</i>	Yes	No
<i>Library</i>	No	Yes
<i>Or</i>	Yes	Yes
<i>And</i>	Yes	Yes

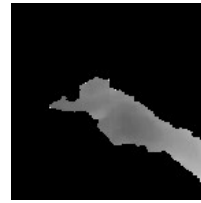
According to Table 1, the five conditions are:

- *None* – where none of the black pixels is ignored on any image,
- *And* – this condition ignores only when both pixels, in the matching, are black,
- *Or* – similar to the *And* condition, the *Or* condition ignores the matching when either pixel is black,
- *Kinect/Library* – skips the matching when the pixel is black, in the respective image, *Kinect* or *Library*.

Through extensive experimentation it is expected to achieve the best condition and, on one hand side, decreasing the computing power and increasing the accuracy of the system, on the other hand side.



(a) Image from Kinect



(b) Image from library

**Fig. 5.** Example of images used in a recognition. Same sign "A"

From the conditions stated before it is possible to see that each condition gives a different output of pixels to compare. To verify how they work, it will be used the images in Fig. 5, where both images are a representation of sign A, to display an example on how each condition works. The Fig. 5a contains the image to be compared, therefore the image that is going to be changed, and the Fig. 5b contains the image that is supposedly from the gesture library displayed.



(a) None

(b) And

(c) Or

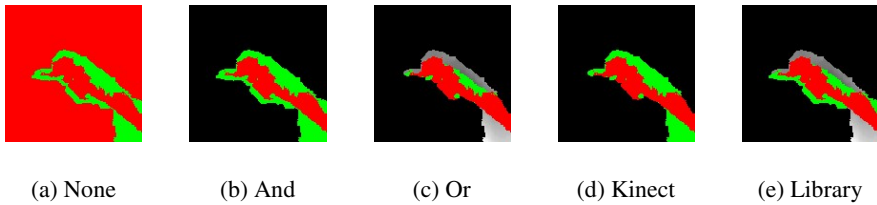
(d) Kinect

(e) Library

**Fig. 6.** Overlay of two images according to the condition, the green pixels shows the pixels used on recognition

In the Fig. 6. , it is shown how the various conditions affect an image, where green represents the pixels available for comparison. With the definition of this conditions for recognition, it's possible to make a matching between two images. This matching gives the accuracy of distances. In other words, when a pixel in the image to recognize is close enough to the equivalent pixel in the library image it flags that pixel as a valid pixel.

After comparing all the pixels, according to a pre-set condition from Table 1, it's determined the accuracy of the matching by dividing the number of valid pixels by the total number of pixels.



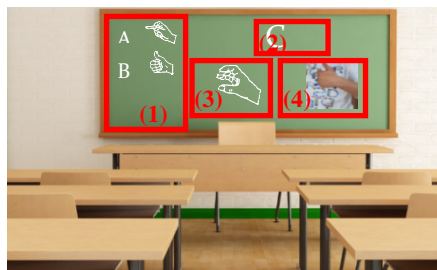
**Fig. 7.** Overlay of valid pixels, in red, according to the distance, over Fig. 6

From Fig. 7, it is possible to see the pixels that gave a positive match, represented in red, when the distance between each figure is smaller than the variance  $V_D$ , in this case the distance, in grey scale, is 25, or in other words, approximately 2cm. In this matching is achieved an 88% approximation rate using the *None* condition and 52% approximation rate using the *Kinect* condition.

## 4.2 Serious Game

One of the main objectives of this work is to build a serious game where the user can learn and play with sign language. Kinect-Sign, the proposed serious game, is divided into two main modes: *School* mode and *Competition* mode, as stated before. The idea behind these modes is to help the families to interact with their hearing impaired loved ones. So, the *School* mode will assist in the teaching of sign language and the *Competition* mode is intended to help strengthen the connection inside the family, done through playful games.

**School Mode.** The *School* mode is designed so that any user can learn sign language in the most easy and playful way, but at the same time in a serious environment. In order to meet these parameters, the environment was designed as a classroom where the user will be enrolled in short lessons. For that reason, the lessons have at most five elements, also so that the user doesn't lose focus while learning.

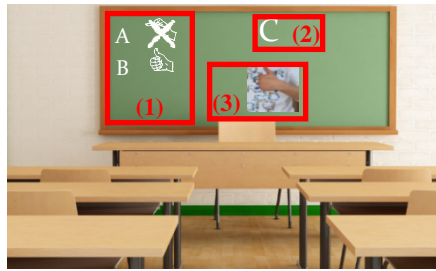


**Fig. 8.** Lesson design: (1) –Elements already learnt; (2) – Element being learnt; (3) – Representation of the element being learnt; (4) – Hand image acquired from the Kinect sensor

The elements learnt in the lessons, as expressed in chapter 1, are letters and so the lessons will be continuous, separated by evaluations after every two lessons. An example of lesson can be viewed in Fig. 8, the user must reproduce with its hands the symbol the game presents and when the image get a good accuracy rate percent sets that element as correct and passes for the next element in the lesson.

After every two lessons, we should verify if the various elements have been learnt. To do so, there were created some evaluations. These evaluations appear whenever there is a need to verify the knowledge of the user. Normally, an evaluation will focus the two previous lessons, taken by the user, and they verify if the user needs or not to remake one certain lesson.

In order to pass an evaluation and unlock the next lesson the user must have at least 60 percent of the evaluation correct, this means that three out of five elements must be right.



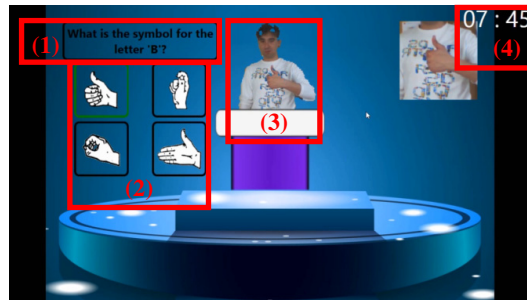
**Fig. 9.** Evaluation design: (1) –Elements answered and if they are correct or not; (2) –Element being evaluated; (3) – Hand image acquired from the Kinect sensor

The displaying of the evaluation is very similar to a lesson, Fig. 8, where the main differences are that the user cannot see the representation of the element he must reproduce, and the elements that the user already answered wrongly will be crossed on the table (1) of Fig. 9.

**Competition Mode.** The Competition mode is the more playful area on this serious game. Therefore it was design so that it may transport the user to a fun environment, while playing games that could relate to sign language. Bearing this in mind, it was developed a TV Show environment.

By designing the TV Show scenario, some games appeared to be more natural, for this type of environment, than others, one example is the *Quiz*. The proposed game is based, as the name implies, on making questions. So, when playing *Quiz*, the user will be asked five questions in order to make the best score possible.

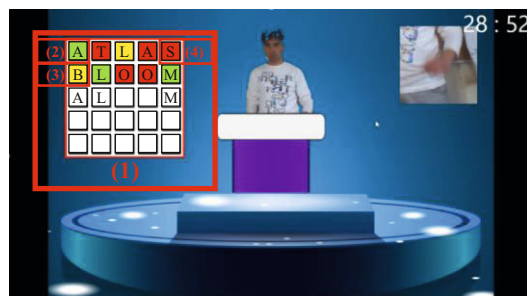
As it's possible to see in the Fig. 10, this game gives four possible answers for the question asked. This is a way to help the user, answering the question, and to help the recognition system, since it's just needs to compare with four elements in the library, making it a faster recognition. Also, in Fig. 10, it's visible that the user has responded to the question correctly, so the border of the answer is green. If the answer was wrong then the border would be red.



**Fig. 10.** Quiz game: (1) - Question made to the user; (2) - Possible answers for the question; (3) – User (4) – Countdown timer to count how much time is left to end the round/game

Other game proposed is the *Lingo*. Very similar to *Mastermind*, this game was created for the user to discover five letter words. Therefore, the idea of the game is for the user to spell, in sign language, the five characters that compose one word and find out if he was right, wrong or if there is correct letters, in the wrong place.

In order to ease the game there is a limit of five tries before giving the word as a wrong word. In the Fig. 11 it's possible to see how this game is going to work.



**Fig. 11.** Lingo Game: (1) - Lingo board; (2) - Right character in the right place; (3) - Right character in the wrong place; (4) - Wrong character

From the figure above, Fig. 11, it's visible the various states for each character. For instance when a character is right and in the right place, Fig. 11 element (2), it's shown in green, when a character is right but in the wrong place, Fig. 11 element (3), it's shown in yellow and if the character does not exist in the word, Fig. 11 element (4), appears in red.

## 5 Validation

The validation chapter is where the explanation on how is the work is tested takes place. In other words, it will be referred what where the experimental work done over the project, including a detailed description on the work, and what were the experimental results obtained from this project. After obtaining all the results, there will be



a brief analysis and discussion on this results, where they will be explained and carefully considered their viability in order to continue the project development.

## 5.1 Sign Language Recognition

SLR is a simplified system where the Kinect sensor acquires an image, with the depth sensor, transforms that image, so that it may return only the right hand of the user in a grey scale, and then uses that image to compare with the gesture library, according to the condition and the distance between images. To validate the SLR system, there are two types of validations where the results cooperate to find the best values for matching.

The first type of validation is the validation through images, in other words, one image is matched with the entire gesture library, using all the conditions and distances (1 to 50 in grayscale). From this validation it is determined the three best conditions and five best distances to use apply in the algorithm. The second type of validation is made through videos, simulating a real-time recognition. During this validation it is going to be used the three conditions and five distances determined on the first validation and it is obtained the condition and distance that will be used in the serious game.

Prior to the validation process it is the constructed a strong dataset to support the recognition system. Therefore, it is acquired 300 images per character, from just one person. Also, from the same person, it is constructed the validation dataset, composed, per character, of six images and three videos, of 300 frames each.

There are two types of values obtained from the SLR validation: 1 – the approximation rate, which refers to the approximation between images, in other words, the ratio between valid and used pixels, as shown in red and green, respectively, in Fig. 7; and 2 – the accuracy rate, or in other words, the ratio of signs acquired for each character.

**Image Recognition.** Image recognition is the process where one image is compared with the entire gesture library and from that values it is possible to extrapolate the three best conditions and five best distances. To implement this recognition it is used the five conditions, described in chapter 4.1: *None, And, Or, Kinect* and *Library*; and distances between images ranging from 1 to 50 in grayscale, equivalent of 0,08 cm to 3,92cm.

As an example, it is shown the results acquired when making the recognition, only with the *None* condition, for the six images that contains the letter 'L', this images can be seen in Fig. 12.

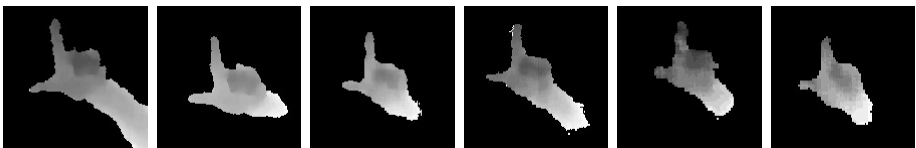


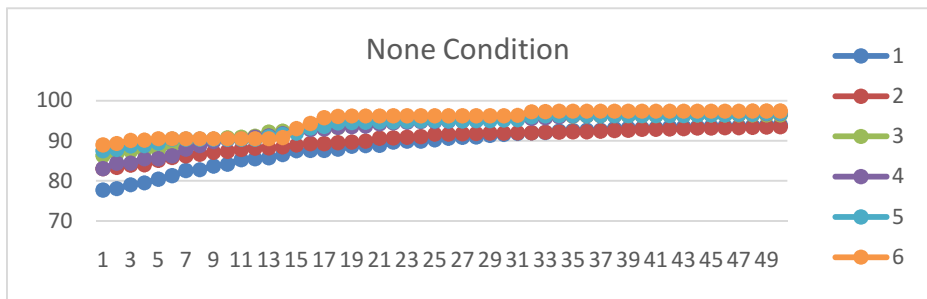
Fig. 12. Images of 'L' to be compared with the gesture library

*None Condition.* As expressed before, the *None* condition is where none of the pixels are ignored when matching two images. Therefore, it is expected to be achieved great values of approximation, in one hand, and, in the other hand, a better accuracy rate.

**Table 2.** Signs acquired using the *None* condition in a sign ‘L’

	G	L	Q	R	U	X
1	10	34	1	5	0	0
2	0	46	0	0	4	0
3	0	50	0	0	0	0
4	0	50	0	0	0	0
5	0	47	0	0	0	3
6	0	48	0	0	0	2

Table 2 shows the dispersion of acquired signs. From the obtained values is possible to conclude that the images used during the recognition are good ‘L’ images, since most of the results are correctly determined.



**Fig. 13.** Graphical representation of the approximation rates

On the other hand, Fig. 13 shows the approximation rates obtained for each image according to the distance used in the recognition. Despite some of these rates not belonging to the wanted sign, it is fairly possible to verify that the rates don’t suffer great variations when the recognition distance is increased. In the *None* condition this is explained by the fact that there is no ignored pixel and most of those pixels are black.

**Global Results.** Now, that it has been seen how the recognition works and the results acquired using the *None* condition, it is time to make the validation for the entire test set and obtain the global results for the image recognition. With these results it will be possible to extrapolate the distances and conditions in which the algorithm works best, as required for the video recognition.

In order to obtain the three conditions used on the video recognition, the accuracy rates are used. This is due to the fact that, as expected, the images recognized have a great tendency to refer to it’ own sign. This accuracy rates are shown in Fig. 14.

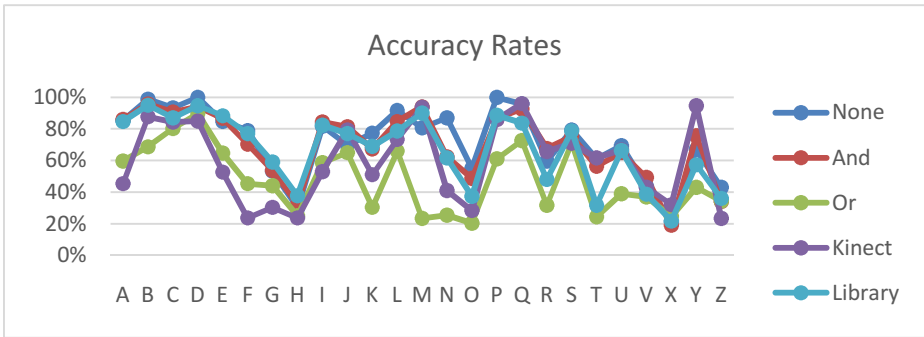


Fig. 14. Average of accuracy rates for all the letters

From Fig. 14 it is clearly visible that the *Or* condition is the worst condition in terms of accuracy rates. Other key element to be retained, from Fig. 14, is the fact that not all the letters have good accuracy rates, this is explained for the fact that not all the letters possess “good” images for recognition and some signs are very similar between images, for instance the signs ‘A’, ‘K’ and ‘T’.

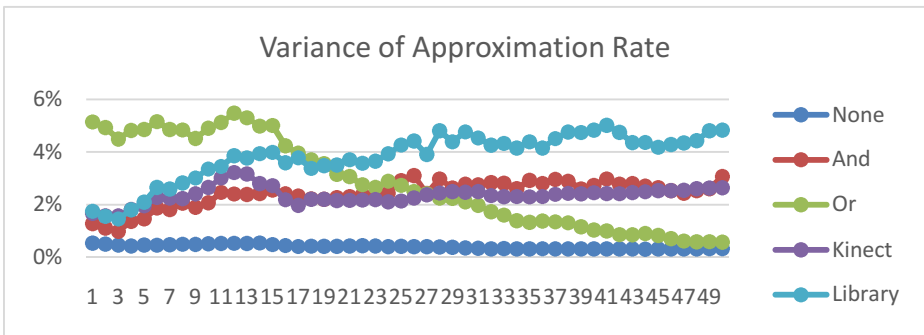


Fig. 15. Average of the variance on approximation rates

The figure above, Fig. 15, is the graphical representation of the average of variances on approximation rates. In other words, as seen in Fig. 15, not all the acquired signs represent the correct sign, so the variance of approximation is the difference between the approximation rate acquired when matching with the entire library and the approximation rate acquired when matching just with the correct sign.

*Single Recognition Conclusions.* From the values obtained in this validation it was decided to use the *None*, *And* and *Library* conditions on the video recognition, since they are the best conditions when it is refer to accuracy results. On the other hand, the values for distance are: 10, 16, 19, 27 and 39. This values where obtained through the approximation rates and the variances from the approximation rate of the acquired sign with the approximation rate of the correct sign.

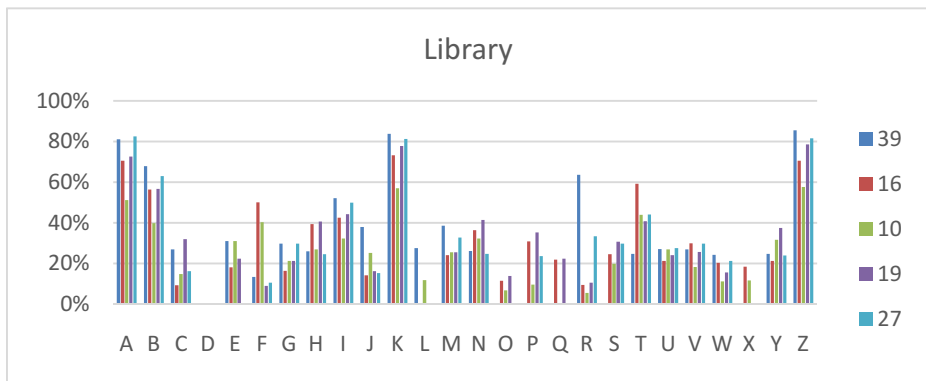
**Video Recognition.** Video recognition is the process where a group of images, acquired in a consecutive way is compared, one by one, with the entire gesture library, determining the best match for each frame of the video. Just like in the single recognition, in this process it will be experimented all the five conditions, but the distances used will be just a small group of five values, determined from the single recognition process. For this process of validation it will be used, as expressed before, three videos per character to make the validation.

*Single Letter.* As written many times, a video is a simulation of a real-time recognition where there is a sequence of 300 images ready to be recognized and those images will be used to validate the recognition system. So, by passing the video through the recognition system it will be possible to acquire the values of Table 3.

**Table 3.** Acquired signs using the *Library* condition on all the videos for sign ‘K’

	A	B	E	F	I	J	K	L	M	R	T	U	V	W	Y	Z
39	106	129	0	1	25	3	507	7	8	20	30	5	4	4	19	21
16	108	126	0	6	24	5	495	0	7	5	58	4	4	5	18	27
10	104	110	9	9	23	4	452	3	7	5	100	9	8	4	14	18
19	110	96	1	3	21	3	570	0	8	5	20	7	3	7	20	21
27	95	160	1	2	35	4	486	0	10	6	41	8	2	3	20	21

This table shows that the sign ‘K’ appears mostly when the set distance is 19, using the *Library* condition. While the distance 10 is a worst distance for detecting ‘K’. Despite this, from the 900 frames, where some are noise, in all the distances more than 50% is detected as the correct sign.



**Fig. 16.** Average of accuracy rates using the *Library* condition

Fig. 16 shows the average accuracies for all the video frames when using the *Library* condition. From it, is visible that, for most distances, the difference between the wanted and the acquired sign is very short. Despite that, is also visible that the sign ‘A’, ‘K’ and ‘Z’ get the higher accuracies,

*All Videos.* Through the validation of all the videos shows what are the best parameters for recognition. Through the video validation it was possible to conclude that the best condition for recognition is the *Library* and the best distance is 19. This parameters are then to be used on the validation of the serious game prototype.

## 5.2 Serious Game

The validation on the serious game resulted of the testing of multiple users of the game. An empirical method was applied on the test subjects of two classes from the “Instituto Jacob Rodrigues Pereira” and on the participants of the eNTERFACE workshop. This resulted on three different test groups: a beginner, with no experience on sign language, a medium, where the group already knew something about sign language and an experienced group, the hearing impaired used to communicate with sign language.

The results were satisfactory, since it was possible to view the enthusiasm of the different participants, and the assistance provided by them in order to improve the current prototype. Still, a more in-depth analysis should be done (e.g. with questionnaires) to validate the developed prototype.

Nevertheless, the gesture library was reduced from 300 images per character to three images per character in order to the recognition process into a real-time process. This allowed the recognition of 4 frames per second. Also, the images used on the recognition where the images that appeared most in the validation process.

For this SLR algorithm it was verified that the serious game prototype works much better when the player and the creator of the gesture library are the same.

## 6 Concluding

In this chapter, it’s pretended to conclude about the results obtained through the experimentation of the SLR system and the usage of this system in the sign language game and, therefore, testing its viability. Afterwards, there will be described on how the eNTERFACE workshop contributed for the development of this project and there will be presented a set of options for future development, subjected to the themes presented on this project.

### 6.1 General Conclusions

The existing documentation is rich in studies on SLR, including studies using the Kinect sensor. Despite this amount of studies, they lack, as much of the researchers’ work, the step of bringing it into market. Therefore, the proposal of this project is to create a serious game, combining sign language and the Kinect sensor, which will lead to a product market-ready.

As seen throughout this report, this game is composed of two parts: the creation of a SLR system and the application of this system on a Kinect based serious game. For this article only the SLR algorithm was subjected only to a more extensive test. To

validate the algorithm, tests were made evaluating the recognition rate and accuracy, as they will be detailed in the following sub-chapters.

**Sign Language Recognition.** In this system it was extensively studied the algorithm responsible for the workings of SLR. There were studied different methods of working and different parameters in order to acquire the best algorithm possible. From validation it was concluded that the *Library* condition and the distance 19 are the best parameters to work with this system.

**Serious Game.** The serious game has suffered a very simple validation, since it was just studied in terms of satisfying and pleasure of playing by multiple users. But, despite the simple study made it was possible to achieve very important conclusions about this project. The most important conclusion is that this is a viable project in terms of recognition system and really it’s an enjoyable serious game.

## 6.2 eNTERFACE Contribution

The work presented in this project has been developed to serve as a master thesis and the passage through the eNTERFACE workshop was a step towards the conclusion of this thesis. This workshop served as a platform to discuss new ideas, also to revisit old ones and to understand features, about the project at hands, which were not possible to understand until then.

Clearly, one lesson that can be returned from this workshop is that every person has different body features. Applied the lesson to the project means that each is unique and that creates the necessity to implement some form of generic module capable of making the SLR. Since one user library will probably not make the right sign recognition for other user.

To conclude, the contribution provided by the eNTERFACE workshop was of great value for the improvement and expansion of the project Kinect-Sign.

## 6.3 Future Works

From the work developed for this project, it’s possible to express that this game is not yet market ready. This means that there is still some work to be done, especially in the SLR system. Therefore, it is required the improvement of the SLR system, which might include machine learning or image processing algorithms. Other improvement is the serious game, particularly, the game UI which can be transformed into a 3D environment, for example using the Unity 4 game engine. Also, for the serious game, an increase in number of lessons and existing games can come to place.

The usage of distinct inputs towards improving the reliability of this teaching approach will also tackle the research area of Collaborative Networks, both in terms of support infrastructures, as mentioned in [18], and in terms of teaching approaches, as for example putting the students to work collaboratively, as mentioned in [19].

Some other important work is required to continue the game development. First of all, a questionnaire about the serious game, for validation purposes. Second, a market

study, to detect the needs of the target audience. And for last, the expansion of the gesture library so that the SLR algorithm might work more accurately.

## References

- [1] World Health Organization, Deafness and hearing loss. World Health Organization (February 2013), <http://www.who.int/mediacentre/factsheets/fs300/en/> (accessed October 2, 2013)
- [2] Cooper, H., Holt, B., Bowden, R.: Sign Language Recognition. In: Visual Analysis of Humans, pp. 539–562. Springer (2011)
- [3] Song, Y., Yin, Y.: Sign Language Recognition (2013), <http://people.csail.mit.edu/yinyin/resources/doc/projects/6867term-project.pdf>
- [4] “Kinect,” Wikipedia (August 13, 2013), <http://en.wikipedia.org/wiki/Kinect> (accessed August 22, 2013)
- [5] Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., Zhou, M.: Sign Language Recognition and Translation with Kinect, Beijing, China (2013)
- [6] Gameiro, J., Cardoso, T., Rybarczyk, Y.: Kinect-Sign - Teaching sign language to ‘listeners’ through a game. In: Conference on Electronics, Telecommunications and Computers, Lisboa, Portugal (2013)
- [7] “Sign Language,” Wikipedia (August 16, 2013), [http://en.wikipedia.org/wiki/Sign\\_language](http://en.wikipedia.org/wiki/Sign_language) (accessed August 23, 2013)
- [8] Winnie, T., Drennan, A.: Sign Language Bingo (2008), <http://www.superduperinc.com/products/view.aspx?pid=bgo133&s=sign-language-bingo#.UheEIxtwquI> (accessed August 23, 2013)
- [9] Sign-O (2007), <http://www.amazon.com/Sign-ASL-Bing-Game-Deaf/dp/B003AX6VO8>
- [10] “Sign the Alphabet,” Funbrain, <http://www.funbrain.com/signs/> (accessed August 23, 2013)
- [11] Breuer, J.S., Bente, G.: Why so serious? On the relation of serious games and learning. Eludamos. Journal for Computer Game Culture 4, n° Serious Games (2010)
- [12] “Kinect Could Reinvent Educational Gaming,” Winextra (2013), <http://www.winextra.com/tech/opinion/kinect-could-reinvent-educational-gaming/> (accessed August 23, 2013)
- [13] Starner, T.E.: Visual Recognition of American Sign Language Using Hidden Markov Models (1995), <http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA344219> (accessed August 24, 2013)
- [14] “Hidden Markov model,” Wikipedia (August 4, 2013), [http://en.wikipedia.org/wiki/Hidden\\_Markov\\_model](http://en.wikipedia.org/wiki/Hidden_Markov_model) (accessed August 24, 2013)
- [15] Correia, M.M.: Reconhecimento de Elementos da Língua Gestual Portuguesa com Kinect (2013), <http://paginas.fe.up.pt/~ee06160/thesis/> (accessed August 19, 2013)
- [16] “Top 10 Kinect Games,” Squidoo (2013), <http://www.squidoo.com/top-10-kinect-games> (accessed August 22, 2013)

- [17] Zoo Games, Minute To Win It (October 2011),  
<http://kinectplus.com/games/minute-to-win-it/> (accessed August 22, 2013)
- [18] Camarinha-Matos, L., Menzel, K., Cardoso, T.: ICT support infrastructures and interoperability for VOs. Virtual Organisations Cluster–VOSTER WP4 D (2003)
- [19] Klen, E., Cardoso, T., Camarinha-Matos, L.: Teaching Initiatives on Collaborative Networked Organizations. In: 8th CIRP - International Seminar on Manufacturing, Florianópolis-SC, Brazil (2005)



# Hang in There: A Novel Body-Centric Interactive Playground

Robby van Delden<sup>1</sup>, Alejandro Moreno<sup>1</sup>, Carlos Ramos<sup>2</sup>, Gonçalo Carrasco<sup>2</sup>,  
Dennis Reidsma<sup>1</sup>, and Ronald Poppe<sup>1</sup>

<sup>1</sup> Human Media Interaction, University of Twente,  
P.O. Box 217, Enschede, The Netherlands

{r.w.vandelden,a.m.moreno,d.reidsma,r.w.poppe}@utwente.nl

<sup>2</sup> Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,  
2829-516 Caparica, Portugal

{carlosjmramos,gbcarrasco}@gmail.com

**Abstract.** We introduce and evaluate a first version of a novel body-centric playground that aims at increasing bodily exertion and immersion. The concept centers around the player, who is suspended from the ceiling using a rope and climbing harness and stands on a tilted platform. This caused players to assume a body posture that elicits the feeling of flying, which was further enhanced by the flying game that they played. We discuss the choices made in the type of body movements, and how these relate to different aspects such as movement mimicry and exertion. We performed a user study, in which the hanging position was compared to a setting where players stood on the ground. We found no significant differences in the amount of movement and perceived engagement between the two conditions. However, there was a tendency of favoring the hanging position. Moreover, we observed that the placement of game elements affected the movement patterns.

**Keywords:** Body-centric, interactive playground, engagement, exertion, postures.

## 1 Introduction

Computer entertainment taps into a variety of the prerequisites to feel good in life [21]. At the same time, there are growing health concerns because these advances in technology have led children to adopt a sedentary behavior, causing an increase in obesity [22]. Nonetheless, the introduction of the very technology that caused this issue can also be used to counter it, effectively turning the problem into a solution [7]. Researchers have begun to study how to create interactive playgrounds, interactive installations equipped with sensors and actuators where children are encouraged to exercise and interact socially through engaging game experiences [19,20,9]. These interactive playgrounds are co-located interactive spaces, ranging from a few square meters to public squares equipped with a

range of sensors and feedback devices. The advent of affordable, robust technologies such as Microsoft Kinect, Arduino, projections and LEDs greatly facilitates the implementation of such playgrounds.

Studies have shown that making the body the center of the experience in these installations leads to higher engagement and immersion [2] and promotes social interaction [8]. This greatly attributes to the success of games that use bodily interaction, e.g. using Nintendo Wii-motes and Microsoft's Kinect. Often, the gameplay of these games is straightforward. However, the link between engagement and bodily exertion is bi-directional: games that provide an engaging gameplay also encourage people to move more [5,1]. This increases their immersion in the game, thus creating a feedback loop where gameplay, engagement and physical exertion play key roles [1]. Given that these three elements are strongly related, it makes sense to actively consider them in the design of body-centric interactive installations.

Pasch et al. [15] identify four aspects that relate players' movements to their immersion and engagement levels: natural control, mimicry of movement, proprioceptive feedback and physical challenge. Each of these have informed the creation of the first prototype of Hang in There, a novel body-centric interactive playground (See Figures 1 and 2 for an impression). In this paper, we describe and evaluate the playground, which is aimed at promoting physical activity by providing a new user experience that focuses on creating an increased sense of immersion. A virtual 3D world with a digital version of the player is shown on a large screen in front of the player. The aim of the game is to fly through this world and collect as many coins and power-ups, while avoiding obstacles and power-downs. The game is controlled by full-body movements. To provide a sensory stimulus of flying, the player is suspended using a climbing harness and rope while standing on a 45-degrees tilted platform.

We present an overview on postures, physical exertion and engagement in games and simulations in the next section. The design, setup and implementation of the Hang in There installation is discussed in Section 3, followed by a user study to investigate the difference in engagement between playing the game in a standing and hanging position. We conclude with a discussion on our findings and present avenues of future work.

## 2 Exertion, Postures and Engagement

Several studies have shown that body-centric control in games increase the engagement and immersion of the user (see [1,14] for an overview). Furthermore, Riskind and Gotay showed that putting someone in a certain posture can have a effect on their level of task persistence, expectations and (self-)perceived emotional state [17]. Bianchi-Berthouze et al. [2] tested how playing Guitar Hero with a standard PlayStation controller and a guitar-like controller affected the engagement of the player. They found that a more natural way of control encourages the player to immerse himself in the game, which leads to a greater emotional experience. In another study with Guitar Hero they instructed users



**Fig. 1.** Exploration of positions

**Fig. 2.** The final prototype

about the ability to exhibit a more expressive movement and posture, tilting the guitar instead of pushing a button. This led to a different type of engagement and play, which tended to change to a more fantasy rich and more emerged play [2]. Based on these and other studies Bianchi-Berthouze concluded that affording body movements linked to role-play resulted in higher player engagement and presence in the virtual world, users becoming more socially and probably more emotionally involved [1].

Making use of affording posture change has been done in some commercial games, simulators and art installations. There are virtual glider games in which users hang in a glider in order to fly in a virtual environment, such as the Dream Glider by Dreamilty Technologies, “Virtual Hang-Gliding over Rio de Janeiro” by Soares et al. that even included a fan for simulating the wind [18], and another “Virtual hang gliding over Rio” by YDreams for Rock in Rio 2011. There are also several flight simulators that tilt the user, including the Jetpack Simulator by Martin Aircraft Company. In this training simulator a user is strapped on a small tilting platform that rotates slightly above the floor with the center point of rotation around hip height. With respect to sports there are also simulators which allow for realistic postures and movement that make the user more engaged and primarily attempt to train the user efficiently. For instance, Skytechsport created the Alpine Racer downhill simulator, in which the user experiences forces including vibrations and tilt, limited to a flat horizontal plane, that simulate real skies and slopes. Ars Electronica Futurelab created Humphrey (II), an interactive

body-centric flight simulator. The user was hanging in a suit that is suspended from the ceiling with ropes. Furthermore, they implemented a force feedback system to simulate the forces during flight. The user could either fly through the city of Linz or dive in the Danube by making body movements.<sup>1</sup>Fels et al. exhibited an installation that allowed users to swim while being in the air [4]. The participant was suspended in a harness that was suspended from a frame with several ropes. In the 1990s Sega created a rotating arcade cabinet called the R360. The user was strapped onto the cabinet's chair similar as is done in a roller-coaster. It had two rotating axes allowing for full rotation, the users could thus even be upside down. Pasch et al. [15] studied how movement in sports games affected the game experience using the Nintendo Wii. They observed that play styles changed according to the motivation or goal of the players, namely if they were playing to relax or to compete. They also found that control is an important factor in the experience. By observing and interviewing the players, they were able to define four movement features that affect players' immersion: natural control, mimicry of movements, proprioceptive feedback and physical challenge. Natural control takes into account how similar the movements of the player are with respect to those they would perform in real life. Mimicry of movement refers to the mapping of the players' movements to those of their avatars or, if there is no avatar, in-game movements. Proprioceptive feedback encompasses the sensory input. Lastly, physical challenge refers to how much exertion the game requires. Two of the main goals of interactive playgrounds are providing engaging, fun experiences and encouraging physical activity [16]. Therefore, these four features are specially important in the design of body-centric interactive installations since they link engagement and immersion to a player's movement and, by consequence, to their physical activity.

Yao et al. [23] built a tangible "RopePlus" installation in which people could remotely interact and play two games: "Multi-Fly" and "Multi-Jump". The main element of the system was a rope connected to a motor which sensed and limited the movement of the rope. In its handle, the rope had an accelerometer and a wireless communicator. When any of the two games was projected, the player could either fly a kite or play jump the rope by using the rope that was provided by the system. This is a typical example of natural control, as the movement are similar as those performed with a real kite, or when jumping rope on a traditional playground. Another example is the "Kick Ass Kung Fu" installation designed by Hämäläinen et al. [6]. They designed a martial arts game where a player could fight AI or player-controlled virtual characters. The setup was composed of a cushioned arena, with one big screen located at each side, where the player was allowed to move. A webcam was used to observe the player who was, after some computer vision processing, inserted into the game as one of the fighters. On these screens, a standard 2D fighting platform game was projected, and the player could move from side to side while punching, kicking, jumping or doing

---

<sup>1</sup> A more detailed description of the Humprey II can be found on [http://90.146.8.18/en/archives/picture\\_ausgabe\\_03\\_new.asp?iAreaID=294&showAreaID=294&iImageID=44618](http://90.146.8.18/en/archives/picture_ausgabe_03_new.asp?iAreaID=294&showAreaID=294&iImageID=44618)

acrobatic moves to attack the enemy. Again, these movements are a natural way of control, and they are mimicked in the virtual world. As the movements were exaggerated in the virtual world, the player was persuaded to make even more spectacular jumps and kicks.

Recently, a lot of attention has been given to installations that focus on the physical challenge feature. Mueller introduced the term exertion interfaces for games that require the player to undergo physical exertion [10]. He argued that this component makes the game more fun and encourages social bonding. For instance, Mueller et al. [12] designed a game of table tennis for three participants. The setup was similar to that which one uses when playing alone, i.e. instead of a net, a wall is located in the center of the table so the ball bounces back to the player when he volleys. The other players were back-projected to this wall, each one occupying half of the wall. The players competed against each other by breaking blocks that were overlaid on the video feed. They found the participants engaged in social interactions often and enjoyed the competition element, even though they sometimes got exhausted.

Although exertion games are able to increase a player's engagement and immersion levels, there is also an important relationship between gameplay and the willingness to exercise. In their "Astrojumper" game, Finkelstein et al. [5] found that gameplay attractiveness was related to exercise effectiveness. They developed a virtual reality game where players had to dodge virtual planets that sped towards them. The players wore stereoscopic goggles along with trackers on their bodies to track their movements. These trackers gave the players a lot of freedom in their movements, as they could move to the sides, jump, duck, and shoot lasers to destroy objects. At the end of the playing session, the participants stated they enjoyed the game and were extremely motivated to play again, demonstrating that attractive gameplay can encourage physical activity. Similar observations were made for the "Hanging off a Bar" installation [13]. The setup consisted of an exercise bar that players were to hang onto while a game was projected underneath them. The game consists of a flowing river where, occasionally, a raft drifts down and the player is allowed to let go of the bar. However, once the raft drifts off the player has to hang onto the bar again. The player loses if he lets go of the bar and "falls" into the water. In the evaluation phase, players mentioned that they would invest more physically if, among other things, the game facilitated fearful emotions or allowed to play with other people, both of which make up elements of gameplay.

### 3 Hang in There System Design

The different movement features and their relation to engagement must be considered when designing body-centric interfaces. In this section, we discuss how we designed an engaging installation that supports proprioceptive feedback, natural control, movement mimicking and physical exertion, at the same time. By carefully designing the different game elements, we aim at increasing immersion and engagement, with the ultimate goal of increasing the fun experience and the amount of movement.



**Fig. 3.** Screenshot of gameplay with avatar in starting position

### 3.1 Physical Setup

We set out to design a novel body-centric interactive playground that elicits the experience of flying in the user. This feeling is created by putting players in a extreme forward tilted position. As a consequence, the center of gravity is in front of the feet, forcing the body to fall forward. To accomplish this, the player stands on a smooth, tilted platform while being attached to a climbing rope and harness (Figure 2). Different tilt angles for the platform were tested (Figure 1) to experience the feeling of suspension, and the comfort of the position and equipment. Being totally suspended provided a great experience, however it was unfeasible to maintain this position for long because of the pain caused by the chosen equipment. This could be remedied by adding more support points at the cost of a reduced feeling of freedom. Even with the current setup, there is a certain level of discomfort, especially around the hips and abdomen, that limits the maximum amount of gameplay. Better/different equipment could be tried in the future.

### 3.2 System Architecture

We had a single system consisting of a game engine with a Kinect for Xbox sensor running on a single laptop, outputted to a projector and speakers. The game is created in Unity 4.1, a free to use 3D game development system and engine. To

interpret and deal with the movements of the user, we make use of Kinect for Windows SDK 1.6 and an existing Kinect wrapper from the asset store of Unity 3D<sup>2</sup>. To allow for reasonable performance and ease of use we altered some parts of this wrapper. We removed the automatic rotation of the Kinect sensor and the gesture recognition functionality.

### 3.3 Game Controls

The game consists of a player flying through a virtual world, in which he is represented by a 3D avatar in third-person view (see Figure 3). The world is projected on a wall in front of the player. Once a player is recognized with the Kinect skeleton tracker, the avatar will move forward at a constant speed. The game lasts 109 seconds, and ends with the player going through the finish line, since the speed of the player is unaffected by any game interaction.

The player's movements are mimicked by the avatar and are used to navigate through the virtual world (see Figure 4). Specifically, the player can navigate to the left or right of the virtual world by moving to his left or right respectively. This movement is physically demanding because his center of gravity is in front of him, but also because the player is drawn back to the center because of the harness attachment point. Therefore, maintaining a steady position on any one side requires the player to use muscles located in the legs and abdomen. The player can also ascend, descend or maintain his current altitude in the virtual world by quickly flapping his arms, holding them close to his body, or extending them outwards respectively. These movements exercise the arms and shoulders.

The Kinect is placed on the floor pointing towards the person hanging on the platform. In this set-up, the legs were not always recognized properly, among other reasons due to the short distance between the player and the Kinect. Therefore, we decided to remove the bindings of those "Kinect joints" to the avatar. Instead, the legs simply followed the rotation of the hip bone, always fully extended.

We evidenced that the instinctive way to flap the arms changed from player to player, but it also depended on their current position. When hanging, most players moved their arms up and down along their body, while others moved them back and forth, as if they were birds. In the standing position, almost every player moved their arms up and down along their bodies. Since the platform where players stand on is tilted approximately 45 degrees, we allowed users to flap either up and down, or back and forth. The elbow joints gave the best recognition results for these movements. Hands, for instance, were sometimes not recognized if they were too close to the body or too far to the sides for the Kinect to recognize.

The recognition of flapping is done with the displacement in y-axis (vertical) and z-axis (depth perpendicular to Kinect) of the elbow joints. This displacement is normalized with the length of the hips to the shoulders, to even out differences

---

<sup>2</sup> *Kinect with MS-SDK*, version 1.3, <https://www.assetstore.unity3d.com/#/content/7747>

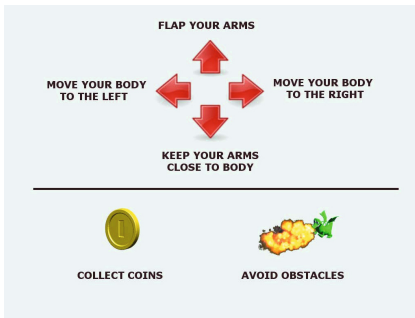


Fig. 4. Game instructions

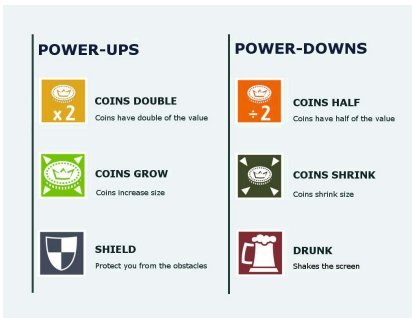


Fig. 5. Power-ups and downs

in body lengths between subjects. If the sum of displacement over the last 12 frames (approximately 0,5 second) is higher than a predetermined threshold, the avatar moves up with a force proportional to this displacement. Falling, or descending, occurs when the player keeps his arms close to the body. The vertical distance between the elbows and hips are normalized based on the distance between the hips and shoulders. If this normalized distance is below a certain threshold, the avatar will descend with a force proportional to this value. These forces should allow the user to have some control over the vertical movement.

Even though we wanted the players to use their whole body to play, the fact that the Kinect could not always recover the location of the legs was an inconvenience. We circumvented this by looking at the absolute position of the player's hip joint in space. By stepping left and right, this joint moves and we use this to position the player's avatar. We tested several multiplication values to empirically find one that allowed the users to move comfortably and fairly, while exaggerating the movements slightly to give them a sense of empowerment.

### 3.4 Game Objects

The goal of the game is to collect as many points as possible by collecting coins that float in the air. However, there are obstacles represented as fire spitting dragons<sup>3</sup>, that will reduce the number of points. To persuade the players to move more to the sides, which should require more physical effort, more coins are placed to the sides of the virtual world and more dragons in the middle. In fact, we placed more coins at the right side of the screen, compared to the left side, to be able to measure whether coin placement would affect the players' sideways movements.

Furthermore, there are several power-ups that make it easier to collect points and power-downs that make it harder to do so (see Figure 5). These modifiers, which are represented as textured boxes, are placed throughout the virtual world to make the game more engaging and add some variation into the gameplay.

<sup>3</sup> Dragon model *Larry* obtained from <http://www.sharecg.com/v/37318/>



There are three classes of modifiers in overall, with each class having a power-up and a power-down. The first class affects the amount of coins the player collects (“coins double”, “coins half”), the second affects the size of the coins present in the field (“coins grow”, “coins shrink”), and the third affects the context of the game (“shield”, “drunk”). There are 8 instances (4 power-ups and 4 power-downs) of each class scattered in the game, e.g. 4 modifiers to double the collected coins and 4 modifiers to halve the collected coins. The only exception is the contextual modifier class, which had 8 power-downs and 4 power-ups. The first two classes are self-explanatory, however the contextual modifiers warrant an explanation. The “drunk” power-down severely shakes the screen of the player when collided against, making it harder to collect coins for a brief period of time. The number of instances of this power-down was doubled because it had a high entertaining value during the development of the installation. The other contextual modifier, the “shield” power-up, prevents the player from losing coins when colliding against dragons (or their fire) once. The duration of the shield is short, and disappears if not used. The shield is represented as a purple, semi-transparent half-sphere that directly in front of the player.

Multimodal feedback is provided during the game when different events occur. The player’s score is shown in the top center of the screen to encourage him to collect as many coins as possible. Encouraging messages, along with appropriate sounds, are also triggered when players collect power ups. When hitting the dragon or its fire, the camera shakes and a short burning sound is played. Every collected coin triggers a “Mario Bros” coin collecting sound. When the avatar of the user is flying too low or too high, a warning sign is shown. When the players reaches certain milestone scores, he is praised with “*Awesome!*”, “*Great!*” or “*Keep going!*” messages. Upon crossing the finish line, a firework sequence is shown and a cheering sound is played.

## 4 User Study

We wanted to evaluate whether being suspended elicited a sense of immersion and increased the engagement of the Hang in There installation. To do this, we tested two different conditions in which the players played the game while standing or in the tilted hanging position. The design, physical setup, procedure and results are discussed in this section.

### 4.1 Experimental Design

In order to compare the standing and suspended conditions, the game and the movements to control the avatar were kept equal. Therefore, only the physical position of the body changed. We used a within-subjects design, so each participant played the same game once in both physical positions. The order was alternated between sessions. There is no standard method to evaluate exertion interfaces [11]. Here, we use questionnaires, an informal interview, and an analysis of the movements of the players.



**Fig. 6.** Standing condition of Hang in There

## 4.2 Physical Setup

We conducted the experiments at the Universidade Nova de Lisboa. The setup was placed in a corridor of the main building of the faculty of science and technology. The physical setup for the experiment consisted of a tilted platform, a harness, a climbing rope, a projector, a laptop, a projection surface and a Kinect. The tilted platform was built with wooden boards that were fixed together and supported by table frames with several counterweights. The slope was about 45 degrees. In the hanging position, the participant stood on the tilted platform, held by the harness. The player faced the screen in a tilted position, assuming a body posture that elicits the feeling of flying. In the standing condition, the participant assumed an upward position, parallel to the screen. During the standing condition, the player was still held with the harness to limit the motion and keep the amplitude of movements from left to right similar in both conditions (Figure 6). Again, the participant faced the screen.

The projector was placed on the floor, behind the platform, by cutting a hole in it. This was done to maximize the projection area, while taking care to avoid casting shadows. A  $3 \times 2$ -meter projection cloth was fixed to a wall, about two meters from the platform. The large dimensions of the screen were thought to

facilitate the feeling of immersion in the game (see Figure 6). The Kinect camera was placed on the floor, just in front of the cloth.

### 4.3 Procedure

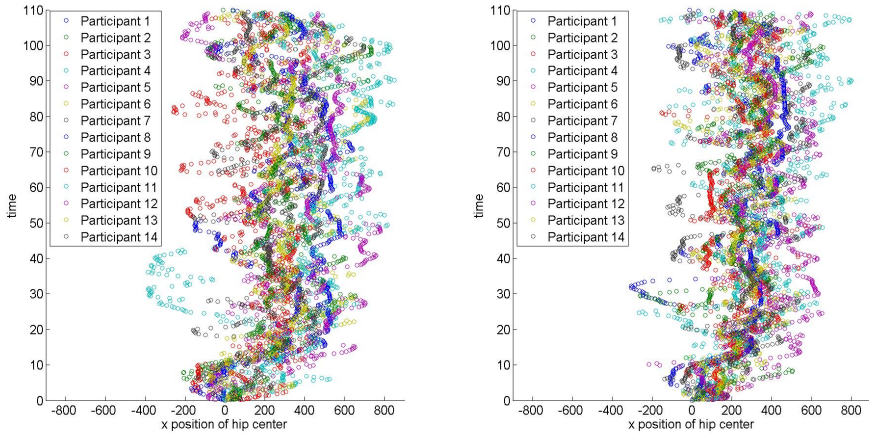
After signing a consent form, the harness was put on and the rules and control of the game were briefly explained. The participant had up to two minutes to test run the game and familiarize himself with the movements and controls. The test run was played while the participant was standing on the floor. Afterwards, the first game was played, in a hanging or standing position, depending on the specified order. Immediately after playing the game, the participant filled in an online version of the Gaming Engagement Questionnaire (GEQ Revised) [3]. In our revised version of the GEQ questionnaire, items that did not relate to the concept of our game were excluded and some of the terminology was adapted to our installation (see Appendix A for the revised questionnaire). Following this, the participant returned to the platform to play again in the second condition. Once finished, the participant again completed the GEQ. Finally, we asked brief questions about the game experience. For each participant, the duration of the entire experiment was around 20 minutes. In addition to the questionnaires and interview, we recorded the joint data of the Kinect skeleton tracker. These joint positions and rotations were saved every 6 frames.

### 4.4 Results

Fourteen participants played the game (11 men, 3 women, average age: 30.2 years, age range: 24-44 years). They were researchers and students of different fields, including computer science and psychology. Each participant played the game twice, once in each condition.

**GEQ-R.** We excluded the question “*How often do you play games from this genre?*” of the revised GEQ as it does not measure engagement. The results for questions 17,18,21 and 22 were mirrored in order to have a positive value corresponding to an increase in engagement for all questions. We first checked the correlation between the items, which resulted in a Cronbach’s alpha of .843 if both sessions are included. This suggests that the questionnaire successfully measures a scale for engagement. Although the mean engagement value for the GEQ-r was slightly higher for the hanging ( $\mu = 4.91, \sigma = 0.65$ ) compared to the standing position ( $\mu = 4.77, \sigma = 0.64$ ) a paired-sample two-tailed t-test did not show significant difference between the two conditions.

Next, we investigate some questions into more detail. The majority of the players ( $n = 8$ ) answered they were more inclined to play the game again for the hanging condition. The other six participants gave the same response for both conditions ( $\mu = 5.79, \sigma = 0.97$  for hanging and  $\mu = 4.79, \sigma = 1.53$  for standing). A paired sample t-test indicated a significant difference ( $p < 0.01$ ). When corrected for the number of questions, to eliminate the probability of obtaining significant differences by chance, we divided  $\alpha$  by the number of questions



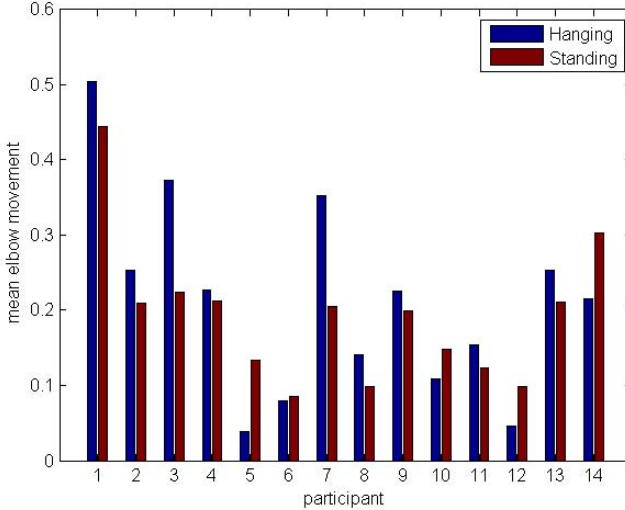
**Fig. 7.** Scatter plot for standing (left) and hanging condition (right), representing hip position on the x-axis in mm

and found that the difference was not significant. We found marginal differences for questions Q2 (“How much delay did you experience”) ( $p = 0.06$ ) and Q20 (“How completely were you engaged in the game”) ( $p = 0.07$ ), indicating a favorable trend towards the hanging condition. Other questions did not approach significance level.

**Movement Logs.** Next, we turn to the tracked joint positions. We noticed that the position of the extremities (hands and feet) were not always estimated. In the game, we therefore used the position of the elbows to detect the flapping. Moreover, we did not use the positions of the knees and feet, but rather the estimated hip positions. In this analysis, we take a similar approach.

We start by analyzing the variation in the hip center in the direction parallel to the screen. Overall, results showed no difference between both conditions in terms of movement in this direction. Since one of our goals was to persuade people to move to the sides instead of staying in the center, we analyzed the position of the avatar. To do this, we needed to compensate for the start position, since the data showed that people started, on average, 2 cm (standing) and 4cm (hanging) to the right. The average position of the players was significantly more to the sides than the center, as measured with a one-sample two-sided t-test ( $p < 0.001$ ), with the average hip position on the x-axis over both conditions per participant. We found a systematic bias to the right of 28cm on average. This tendency to move to the right corresponds to the distribution of the coins, as more coins were located on the right side of the field. The importance of this finding is discussed in the next section. See also Figure 7.

Next, we turn to the vertical movement in the game, which was due to the flapping and holding the elbows close to the body. In Figure 8, we show the movement of elbows. We calculated the distance between the two elbows and



**Fig. 8.** Bar plot showing mean movement of elbow distance in mm over time

took the difference between successfully recognized frames. This is directly related to the actual recognized elbow joints and to the intention to flap. There is no significant difference in the mean amount of movement between the two conditions.

**Open-ended Interview.** Five participants clearly indicated that they preferred the hanging over the standing condition. For instance, one participant stated *“Hanging was much more fun, the other one is not so special”* and some of them stated they wanted to play the hanging game again, or that the session was too short. The other participants were not as open about their preference. Interestingly, even though the players incurred physical discomfort in the hanging position, no one indicated they preferred the standing position. Apart from their preference, four participants indicated the control or recognition had to be improved, e.g. *“The translation is a little bit unnatural, was expecting to be more like a child flying”*. This was due to the fact that instead of stepping to the sides to make the avatar go to any one side, he thought it would be better to rotate his upper body to do so.

## 5 Discussion and Future Work

We have implemented Hang in There, a novel body-centric installation that aims to encourage physical exertion and an enhanced sense of immersion in the player. To this end, we designed an installation where the user is suspended on a climbing rope while playing a coin-collecting game. The player’s avatar is controlled using body movements. We wanted to research how body position and movement could

affect the engagement of the player. The interactive playground was tested with fourteen participants that each played the game in two conditions: in a normal standing position and hanging in a forward tilted position.

Our results do not show a statistical significant difference in engagement between the two conditions, although there appeared to be a tendency to be more engaged in the hanging condition. Testing more participants could prove whether this difference is significant. Several participants indicated they liked the experience of hanging more. When asked if they would like to play the game again, they were more inclined to play in the hanging position, approaching significance on a corrected confidence level. The hanging condition did not lead to an increased amount of body movement. In fact, there was less arm movement in the hanging condition. One possible explanation of this observation is that keeping the body upright in the hanging condition required exertion. With the current measurements, we are not able to investigate this into more detail. Measuring muscle activations might provide the data to perform these analyses.

An important finding came from the displacement in the horizontal direction of the players. Since we wanted the players to move sideways instead of staying in the center of the field, we located the majority of the coins on the sides. Moreover, there were more coins to the right of the field than to the left. We found that, in both conditions, the players were moving to the right more often than to the left. This shows that it would be possible to modify how people play a game, or persuade them to play in a specific manner, by changing the elements of the game and their disposition. Instead of scattering the coins all over the field, in a future version we might place them in other patterns, to stimulate the players to follow these patterns in their movement. Having a trail of coins that crosses the field might be better, since the players would have to move more in order to score.

Based on the positive feedback of the participants, we think our installation has potential for further development. The action recognition, control of the game and comfort could be improved, to increase the enjoyment of playing the game. To reduce the delay between body movement and in-game response, induced both by the Kinect recognition software and the game's working principles, we will consider more direct measurements. One solution could be to use a Wii-controller or accelerometer attached to the upper arm for recognition of flapping. A solution that is considered for detecting movements to the sides, is the use of less demanding computer vision algorithms based on the depth channel of the Kinect. Furthermore, new versions of the Kinect sensor might also lead to a reduction of this delay.

Given the close relation between engagement and social play, we expect that the game can be more enjoyable, but also lead to more exertion, when played with two players. Future work will be aimed at investigating how competition and cooperation can be achieved, and how this impacts both the engagement and level of exertion. To this end, two players could be suspended from a rope, side-by-side. Even though each player has their own side of the platform, interesting interactions could take place in the center of the platform. For instance, players

holding hands to remain in the center to avoid obstacles, or pushing each other to the sides when competing for resources in this shared physical space. In such a scenario players are not required to fixate solely on the projection to keep performing in the game. These kind of interactions could allow for a further increase of social play [1].

Moreover, we consider improving the immersion of the game by 3D projections in combination with 3D glasses. To further add to an increased sense of flying, future work could include fans, more realistic sound effects and 3D sound localization.

Finally, we expect that the Hang in There installation provides a suitable platform for the research into exertion and engagement. We plan to investigate more closely how certain aspects of the game play such as the control and body position affect the level and type of exertion. Measurements of muscle contraction and more accurate body movement could help in shining a light on this topic.

**Acknowledgments.** This publication was supported by the Dutch national program COMMIT. The authors would like to thank Yves Rybarczyk and Tiago Cardoso for organizing eINTERFACE 2013, of which the project “Body-Centric Play” was part of. We would also like to thank Dirk Heylen for his support and Jan Kolkmeijer for sharing his expertise of Unity 3D.

## References

1. Bianchi-Berthouze, N.: Understanding the role of body movement in player engagement. *Human-Computer Interaction* 28(1), 40–75 (2013)
2. Bianchi-Berthouze, N., Kim, W.W., Patel, D.: Does body movement engage you more in digital game play? and why? In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) *ACII 2007*. LNCS, vol. 4738, pp. 102–113. Springer, Heidelberg (2007)
3. Chen, M., Kolko, B.E., Cuddihy, E., Medina, E.: Modeling but NOT measuring engagement in computer games. In: *Proceedings of the International Conference on Games + Learning + Society Conference*, pp. 55–63. Madison, WI (2011)
4. Fels, S., Yohanan, S., Takahashi, S., Kinoshita, Y., Funahashi, K., Takama, Y., Chen, G.T.-P.: User experiences with a virtual swimming interface exhibit. In: Kishino, F., Kitamura, Y., Kato, H., Nagata, N. (eds.) *ICEC 2005*. LNCS, vol. 3711, pp. 433–444. Springer, Heidelberg (2005)
5. Finkelstein, S., Nickel, A., Lipps, Z., Barnes, T., Wartell, Z., Suma, E.A.: Astro-jumper: Motivating exercise with an immer sive virtual reality exergame. *Presence: Teleoperators and Virtual Environments* 20(1), 78–92 (2011)
6. Hämäläinen, P., Ilmonen, T., Höysniemi, J., Lindholm, M., Nykänen, A.: Martial arts in artificial reality. In: *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, Portland, OR, pp. 781–790 (2005)
7. Hillier, A.: Childhood overweight and the built environment: Making technology part of the solution rather than part of the problem. *The Annals of the American Academy of Political and Social Science* 615(1), 56–82 (2008)
8. Lindley, S.E., Le Couteur, J., Berthouze, N.: Stirring up experience through movement in game play: Effects on engagement and social behaviour. In: *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, Florence, Italy, pp. 511–514 (2008)

9. Moreno, A., van Delden, R., Poppe, R., Reidsma, D.: Socially aware interactive playgrounds. *Pervasive Computing* 12(3), 40–47 (2013)
10. Mueller, F., Agamanolis, S., Picard, R.: Exertion interfaces: sports over a distance for social bonding and fun. In: *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, Fort Lauderdale, FL, pp. 561–568 (2003)
11. Mueller, F., Bianchi-Berthouze, N.: Evaluating exertion games. In: *Evaluating User Experience in Games. Human-Computer Interaction Series*, pp. 187–207. Springer (2010)
12. Mueller, F., Gibbs, M.R., Vetere, F.: Design influence on social play in distributed exertion games. In: *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, Boston, MA, pp. 1539–1548 (2009)
13. Mueller, F., Toprak, C., Graether, E., Walmink, W., Bongers, B., van den Hoven, E.: Hanging off a bar. In: *Extended Abstracts of the International Conference on Human Factors in Computing Systems (CHI)*, Austin, TX, pp. 1055–1058 (2012)
14. Nijholt, A., Pasch, M., van Dijk, B., Reidsma, D., Heylen, D.: Observations on Experience and Flow in Movement-Based Interaction. In: *Whole Body Interaction. Human-Computer Interaction Series*, pp. 101–119. Springer (2011)
15. Pasch, M., Bianchi-Berthouze, N., van Dijk, B., Nijholt, A.: Movement-based sports video games: Investigating motivation and gaming experience. *Entertainment Computing* 1(2), 49–61 (2009)
16. Poppe, R., van Delden, R., Moreno, A., Reidsma, D.: Interactive Playgrounds for Children. In: *Playful Interfaces: Interfaces that Invite Social and Physical Interaction*. Springer (to appear)
17. Riskind, J.H., Gotay, C.C.: Physical posture: Could it have regulatory or feedback effects on motivation and emotion? *Motivation and Emotion* 6(3), 273–298 (1982)
18. Soares, L.P., Nomura, L., Cabral, M.C., Nagamura, M., Lopes, R.D., Zuffo, M.K.: Virtual hang-gliding over rio de janeiro. In: *ACM SIGGRAPH 2005 Emerging Technologies*, Los Angeles, California, p. 29 (2005)
19. Soler-Adillon, J., Parés, N.: Interactive slide: An interactive playground to promote physical activity and socialization of children. In: *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, Boston, MA, pp. 2407–2416 (2009)
20. Tetteroo, D., Reidsma, D., van Dijk, E., Nijholt, A.: Design of an interactive playground based on traditional children’s play. In: *Proceedings of the International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)*, Genova, Italy, pp. 129–138 (2011)
21. van Delden, R., Reidsma, D.: Meaning in life as a source of entertainment. In: Reidsma, D., Katayose, H., Nijholt, A. (eds.) *ACE 2013. LNCS*, vol. 8253, pp. 403–414. Springer, Heidelberg (2013)
22. Vandewater, E.A., Shim, M.S., Caplovitz, A.G.: Linking obesity and activity level with children’s television and video game use. *Journal of Adolescence* 27(1), 71–85 (2004)
23. Yao, L., Dasgupta, S., Cheng, N., Spingarn-Koff, J., Rudakevych, O., Ishii, H.: RopePlus: Bridging distances with social and kinesthetic rope games. In: *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, Vancouver, Canada, pp. 223–232 (2011)









# KINterestTV - Towards Non-invasive Measure of User Interest While Watching TV

Julien Leroy<sup>1</sup>, François Rocca<sup>1</sup>, Matei Mancas<sup>1</sup>, Radhwan Ben Madhkour<sup>1</sup>,  
Fabien Grisard<sup>1</sup>, Tomas Kliegr<sup>2</sup>, Jaroslav Kuchar<sup>2,3</sup>, Jakub Vit<sup>4</sup>, Ivan Pirner<sup>4</sup>,  
and Petr Zimmermann<sup>4</sup>

<sup>1</sup> TCTS Lab.

University of Mons, Belgium

<sup>2</sup> Department of Information and Knowledge Engineering

University of Economics, Prague

<sup>3</sup> Web Engineering Group

Faculty of Information Technology, Czech Technical University, Prague

<sup>4</sup> Faculty of Applied Science

University of West Bohemia, Pilsen

{Julien.Leroy,Francois.Rocca,Matei.Mancas,Radhwan.BenMadhkour,

Fabien.Grisard}@umons.ac.be,

{Tomas.Kliegr,Jaroslav.Kuchar}@vse.cz,

{Jaroslav.Kuchar}@it.cvut.cz,

{Vit,Pirner,Zimmermann}@kky.zcu.cz

**Abstract.** Is it possible to determine only by observing the behavior of a user what are his interests for a media? The aim of this project is to develop an application that can detect whether or not a user is viewing a content on the TV and use this information to build the user profile and to make it evolve dynamically. Our approach is based on the use of a 3D sensor to study the movements of a user's head to make an implicit analysis of his behavior. This behavior is synchronized with the TV content (media fragments) and other user interactions (clicks, gestural interaction) to further infer viewer's interest. Our approach is tested during an experiment simulating the attention changes of a user in a scenario involving second screen (tablet) interaction, a behavior that has become common for spectators and a typical source of attention switches.

**Keywords:** user tracking, face detection, face direction, face tracking, visual attention, interest, TV, gesture.

## 1 Introduction

Imagine a 10 years old child in front of his TV today. He will probably be connected to the web, a tablet or a smartphone in his hands, to browse Wikipedia looking for additional information on the animal show that he is looking or sharing on social networks while watching the main screen.

The television, is changing to become more connected, reducing the boundary between the Internet and the television. At the age of high-speed broadband connections, the viewing experience becomes more interactive and connected to extra-content and social networks.

A recent study by Accenture [15] on consumer habits in front of their TV shows that 62% of the viewers simultaneously use a laptop while in 41% of cases a smartphone is used along with the TV. Another interesting point they highlight is that one of the main problems for the content provider is to find how to capture the customer attention by offering the right video content to respond to the viewer expectations. One of the goals of the future TV will be to be able to answer this demand. Most personalization systems are currently based on content personalization by explicit analysis of user actions (remote control, selected channels ...).

In this project and within the research program LinkedTV [1], we are interested in the explicit and implicit analysis of user actions, our goal is to design a tool for personalization based on the analysis of non verbal behavior of the viewer. To do this, the approach we used, is to analyze the attention and actions that a user can have by using a 3D sensor.

The explicit analysis is performed by the integration of both classical remote control interaction and gesture commands.

For the implicit analysis, one of the tracks that we explore is the possibility of detecting viewers' interest during the display of different media fragments on the TV screen. This information is important because it can tell us when, what and how the media interests a user, which will allow to modify the viewer profile without any explicit request or as complementary information along with explicit interactions. To achieve this goal, we implement a solution of head detection and pose estimation using a low-cost depth camera. This choice was made due to the democratization of this type of sensors and their arrival in the home through gaming platforms [24]. Moreover, TV manufacturers begin to integrate cameras into their new systems, regarding the sensors we can see the willingness of the makers to miniaturize sensors such as PrimeSense new camera "Capri" [26]. Thus, we can expect to see in the coming years 3D sensors directly integrated into televisions. But not only integrated into a TV screen, one of the interests of the video connected to the web is its availability across the network on a large number of connected components (smartphone and tablet), in which the 3D miniaturized sensors will soon also be incorporated.

Another aspect that we discussed is the ability to scan the media to assess its ability to attract the attention of a user. We want to measure the level of bottom-up attention within the images. For that a first implementation of an attention mechanism based on rarity was implemented in c++ to enable the rapid processing of a large number of images. Based on this algorithm we proposed the concept of Metadata attention related to areas of media. Useful for understanding the behavior of a viewer may display.

The next section provides information about the technical architecture of the system which will be afterwards detailed as following: section 3 presents the

explicit interaction realized with a gesture recognition method, section 4 focuses on the implicit interactions related to the attention mechanisms while section 5 details the web player and aggregator used to synchronize the viewer behavior and the displayed media fragments. Section 6 presents the experiment we realized to validate our approach and provides some cues about the analysis of the results for media personalization and finally we conclude and present future works.

## 2 Technical Architecture

The system, Figure 1, we have developed has five distinct modules three are integrated into a single workflow and the other 2 are additional elements supporting and adding value to the main module. The three main modules are:

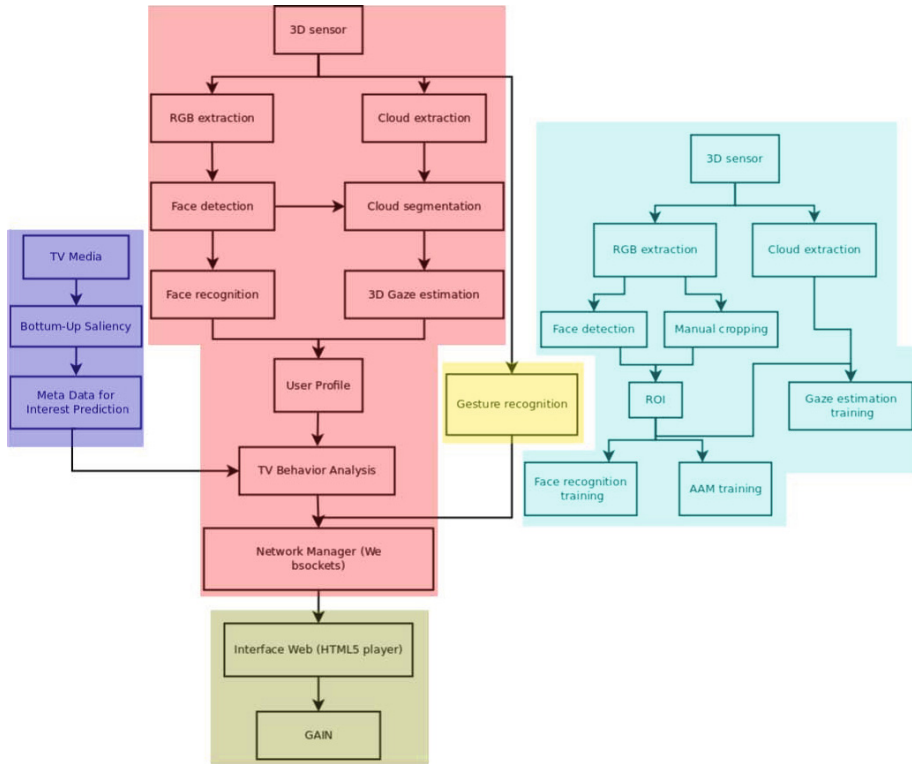
1. Attention Tracker. Implicit analysis module of our system, its goal is to study the movements of a user's head and detect if he looks or not a screen and determines whether he focuses on it. The method used is based on face detection in 2D and 3D then the head pose estimation based on the resulting 3D point cloud.
2. Gesture recognition. Explicit analysis that allows the control of the interface using simple gesture recognition. The recognition technique employed uses the descriptive method by representing a simple gesture like a state machine. The gesture is recognized when all the states are validated.
3. Web interface, HTML5 player and aggregator. This module communicates with other elements using websockets, it is a web player based on Node.js [2], it can play Youtube videos while offering, by analyzing video subtitles, additional content in real time. This module will collect and aggregate all explicit actions taken by the user as well as a status of visual fixation on the player. These data is aggregated to develop a dynamic user's profile.

The other two parallel elements are:

1. Bottom-up attention metadata: the aim here is to augment the media by determining regions of interest that can provoke a bottom-up attention reaction in a subject. The image can be segmented into more or less salient regions. This information is interesting because it gives us a prediction about the probable interest of a viewer, an interest that can then be validated by the attention tracker.
2. Ground truth generator. This additional software is used for annotation and generation of database for face recognition and active appearance modeling.

## 3 Explicit Interaction: Gesture Recognition

There are a lot of possible sensors which allow performing gesture recognition. On one hand, several wearable sensors are available, as accelerometers or gyroscopes. The data they provide does not need a lot of processing before using gesture



**Fig. 1.** The software architecture of the project with the different modules. Red: attention tracker with face recognition and head pose estimation. Yellow: gesture recognition. Blue: attention mechanism and metadata generation. Cyan: ground truth generation. Gray: player and data aggregation.

recognition algorithms, and most of the time, filtering is enough. On the other hand, some sensors use standard or RGBD cameras and provide classical RGB images or/and 3D depth maps and 3D clouds of the scene. In this latter case, the acquired data has to be processed to extract the shape of the user and follow his gestures. For the LinkedTV project, we chose an RGBD sensor. This sensor provides, in addition to classical RGB images a depth map of the scene which describes the objects position related to the one of the camera. An example of depth map is displayed in Figure 2.

The use of an RGBD sensor is also in line with the fact that the same sensor is also used to extract interest information (see section 4). The idea is to use only one RGBD sensor to extract interest and gestures from the viewers. Currently we need to use two different sensors (one for interest and one for gestures) because the head direction extraction needs the current cameras to be not further than 1 meter from the viewer while the context camera needs a global view of the scene and this is thus located on the TV. Nevertheless, the new generation RGBD



**Fig. 2.** RGBD camera depth map. Clear pixels are closer to the camera than dark ones. Post processing on this depth map let us extract the viewer silhouette (in cyan) and the viewer skeleton (red dots linked by red lines).

cameras (like the Microsoft Kinect 2 which will be available in 2014) will allow us to get interest and emotional cues even when the camera is far from the viewer like in the typical living rooms (2-3 meters of distance with the viewers).

For the current developments we used an Asus Xtion depth sensor which is low cost, not wearable, and is able to scan the whole scene. Furthermore, it comes with OpenNI software and the Primesens drivers that allow the find users in the scene and to track their virtual skeletons in 3D (see Figure 2). Among the lots of existing algorithms already used for gesture recognition (Gesture Follower [7], DTW [6], etc.), we chose here the simplest approach: the descriptive method. F. Kistler (from Augsburg University, Germany) developed the Full-Body Interaction Framework (FUBI) [18], an open-source framework that uses a Kinect-like RGBD sensor and that has been successfully used in many situations [20], [19], [16].

The main difference between this method and the others is the learning phase. In other approaches, we have to teach to the system how to interpret the gesture by giving it a basis of examples. With the descriptive method, the user has to learn how to perform the gesture and to do it according to the initial description. The developer defines the gestures either directly in C++ classes or in an XML file (the latter solution being more flexible as modifications can be done and reloaded while the program is running). The gestures consist in some boolean combinations of basic elements, function of the time. These basic elements are



**Table 1.** Gesture implemented and recognized in the project

Referent	Function	Gesture description
Focus	Get the system attention	Draw a circle with one hand in any direction
Play/Pause	Start to play/pause the media	The right hand stay stable 40 cm in front of the torso for at least 2 seconds
Stop	Stop to play the media	Arms crossed for at least 0.5 seconds
Next/Previous	Next/previous media/channel	Right hand moves to right/left quickly
Volume value	Set the volume value	Left arm vertical and right hand near from it, volume = 1 when the right hand is at the same height as left hand and volume = 0 when the right hand is at the same height as left elbow
Mute	Mute the volume	The left hand stay stable 40 cm in front of the torso for at least 2 seconds
Help	Pop up the help menu	Both hands near head
Add bookmark	Add a bookmark on the currently played media	Right hand stay stable 40cm in front of the torso for at least 0.3s and then moves up normally
Remove bookmark	Remove a bookmark on the currently played media	Right hand stay stable 40cm in front of the torso for at least 0.3s and then moves down normally
Lock/Unlock	Pass over controls / accept controls	Left hand above head moves left normally, then left hand above head and moves right normally

the relative position (right hand above head), orientation (left arm oriented front) or linear movements (left hand moves to the left at a minimum speed of 800mm/s) of the skeleton joints. They are updated at each newly acquired frame and give a binary outcome. These binary values are combined in different states, during a defined period of time. All the states make a kind of pipeline, and if the gesture is performed in the order of this pipeline, within the correct timings, it is detected.

For this project, a set of 16 commonly used commands were selected, inspired by Vatavu [29]. According to Wobbrock et al. terminology [31], these commands are called referents. This list of referents should cover all the functions needed to control a TV in a basic use, like navigating into the media, setting the volume, interacting with menus and asking for help. They are presented in Table 1.

We opted for a limited set of referents for two reasons. The first one is the same as proposed by Vatavu [29]: *"The number of gesture commands people can remember for effective use should be limited in order not to increase cognitive load. More individual commands would translate into complex designs with a similar puzzling effect [...] Also, menus represent a viable option to group other, less frequent, commands"* [5]. The second one is linked to the gesture recognition method we use: more gestures could lead to interaction between them and unwanted detections.

To limit the interactions between gestures, we added a "focus" command, a gesture to be performed before most of the other commands to get the attention of the system. If no gestures have been detected after 2.5 seconds, the focus is lost and all the new gestures are ignored until the focus gesture is performed. The TV can be locked or unlocked to prevent any gesture performed in front of the system to be interpreted as a command. It is the same idea as Focus command but in a more restrictive way. Only gestures which need to be done immediately like bookmark do not need to be initiated using the focus gesture.

For flexibility reasons, the gesture description has been implemented in an XML file. Most of the gestures were inspired by [28]. In this experiment, people were told to imagine gestures to match each referent, although the referents were not exactly the same as in our case. After some experiments we agreed on this set of gestures, some of them are used for different referents, depending on the context.

According to FUBI implementation, there are different types of gestures. Postures are static gestures that have to be maintained for a certain period of time to be detected. Linear movement are a simple movement performed at a certain speed (we chose 1m/s for normal speed and 2m/s for fast speed). Combinations are complex gestures which need more than one linear movement to be described. Dynamic postures are like postures but one of the joint is moving and its position, relatively to other joints, is translated into a continuous value (e.g. to control a continuous parameter, such as volume).

As it will be used very often, the focus gesture should be easy to remember and to perform. We chose to implement it as a circle, drawn with the right or the left hand in any direction. The only restriction is to start it from the top.

Each time a gesture is recognized, a message is sent to the attention tracker system (see section 4). The attention tracker packs the message by using the websockets protocol and sends it to the web player (section 5) which is controlled by those gestures. Some controls (play/pause, etc) are fattened by the web player with the video or media fragment ID and time and forwarded to the aggregator system (section 5.2).

## 4 Implicit Interaction: Attention Tracker

Movement and orientation of the head are important non-verbal cues that can convey rich information about a person's behaviour and attention [30][17]. Ideally, to find out if a user is looking at the screen or not, we should extract the ocular movements of the subject. But given the experimental conditions mainly in terms of sensor to viewer distance and in terms of sensor resolution, it is not possible for us to have access to such information. Therefore our system will be based on the assumption that to detect changes in visual focus, the gaze of a person is considered to be similar to the direction of his head. As stated in [25], "[...] *Head pose estimation is intrinsically linked with visual gaze estimation ... By itself, head pose provides a coarse indication of gaze that can be estimated in situations when the eyes of a person are not visible*[...]". Several studies rely

and validate this hypothesis as shown in [3]. Therefore, we will detect visual attention switches and focus by studying the orientation of the head.

Until recently, the literature has mainly focused on the automatic estimation of the poses based on standard images or videos. One of the major issues that must be addressed to obtain a good estimator is to be invariant to variables such as: camera distortions, illumination, face shape and expressions or features (glasses, beard). Many techniques have been developed over the years such as appearance template methods, detector array methods, non linear array methods, manifold regression methods, flexible methods, geometric method, tracking method and hybrid methods. More information on these methods can be found in [25]. More recently, with the arrival of low cost depth sensor, more accurate solutions have emerged [12][10]. Based on the use of depth maps, those methods are able to overcome known problems on 2D images as illumination or low contrast backgrounds. In addition, they greatly simplify the spatial positioning of the head with a global coordinate system directly related to the metric of the analysed scene. Many of these techniques are based on a head tracking method which unfortunately often requires initialization and also undergoes a drift. Another approach, based on the frame to frame analysis as the method developed by [11], provides robust and impressive results. This method is well suited for a living room and TV scenario. It is robust to illumination conditions that can be very variable in this case (dim light, television only source of light, etc.) but is based on a 3D sensor like the Microsoft Kinect.

The approach we propose here is based on the work developed in [23] and [22]. To improve the exploitation and use of our system as an element to be integrated into a set top box, the system architecture and interaction of different elements have been integrated as in shown in Figure.3.

The proposed system is based on the head detection and pose estimation on a depth map. Our goal is to achieve head tracking in real time and estimate the six degrees of freedom (6DOF) of the detected head (spatial coordinates, pitch, yaw and roll). The advantage of a 3D system is that it uses only geometric information on the point cloud and is independent of the illumination issues which can dramatically change in front of a device like a TV. The proposed system can even operate in the dark or in rapidly varying light conditions, which is not possible with face tracking systems working on RGB images. In addition, the use of 3D data provide more stable results than 2D data which can be misled by projections of the 3D world on 2D images. Finally, the use of depth maps let us extract people position and features. This is also important as people detection with no face detection means that the head either has a more than 60 degrees of pitch or 75 degrees of yaw.

The method used here is based on the approach developed in [13][14] and implemented in the PCL library [4]. This solution relies on the use of a random forest [8] extended by a regression step. This allows us to detect faces and their orientations on the depth map. The method consists of a training stage during which we build the random forest and an on-line detection stage where the patches extracted from the current frame are classified using the trained forest.

The training process is done only once and it is not user-dependent. One initial training is enough to handle multiple users without any additional configuration or re-training. This is convenient in a large public setup as the one of people watching TV. The training stage is based on the BIWI dataset [14] containing over 15000 images of 20 people (6 females and 14 males). This dataset covers a large set of head pose ( $\pm 75$  degrees yaw and  $\pm 60$  degrees pitch) and generalizes the detection step.

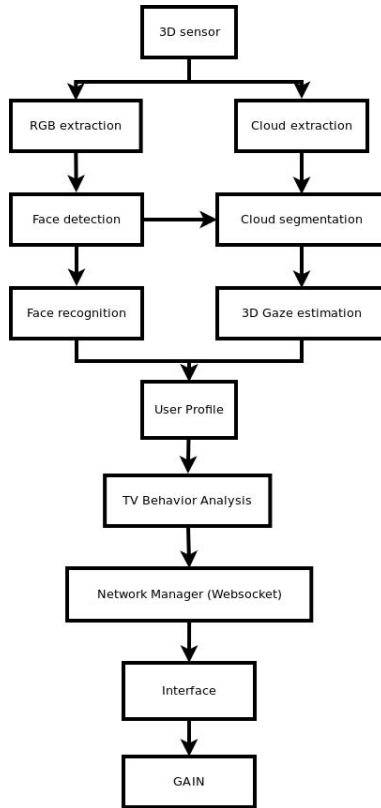
During the test step, a leaf of the trees composing the forest stores the ratio of face patches that arrived to it during training as well as two multi-variate Gaussian distributions voting for the location and orientation of the head. This step of the algorithm provides the head position and a rough head orientation on any new individual without the need of re-training. We then apply a final processing step which consists in registering a generic face cloud over the region corresponding to the estimated position of the head. This last step greatly stabilizes the final head position result.

To improve the performance of tracking and include elements such as face recognition, the major change that we made on the software architecture was to use a 2D face detection (HAAR) as a pre-filter step, Figure 3. This first step performed on the RGB image from the sensor has several advantages:

1. Information limitation. It reduces the cloud information that need to be processed for estimating the users head orientation: the classification of the underlying point cloud can be speeded up.
2. Cross detection. The 2D face detection has also the other advantage to be a predetection test and eliminates some false detections which might occur if the system was only based on the geometrical data.
3. Face recognition. Based on the face detection, we realize a face recognition step to identify the user. This information is used to recognize a known user and to track his beahavior. The face recognition process work by fusing the results of 3 classical face recognition algorithms implemented in the OpenCV library (LBPH, FisherFace, EigenFace).

To detect if a user watches the screen or not, we reconstruct a virtual simplified model of the real scene. Therefore, knowing the 6DOF position of the face of the person detected, it is possible to estimate the point of intersection between the screen virtual model and the orientation of the head (Figure 4). In this way, we can synchronize annotated media with the head tracker and estimate where the user is looking.

This information is sent to a user manager where it is fused with gestural information (obtained as described in section 3) and than forwarded to the network manager module which sends it using websockets protocol to the web player (section 5).

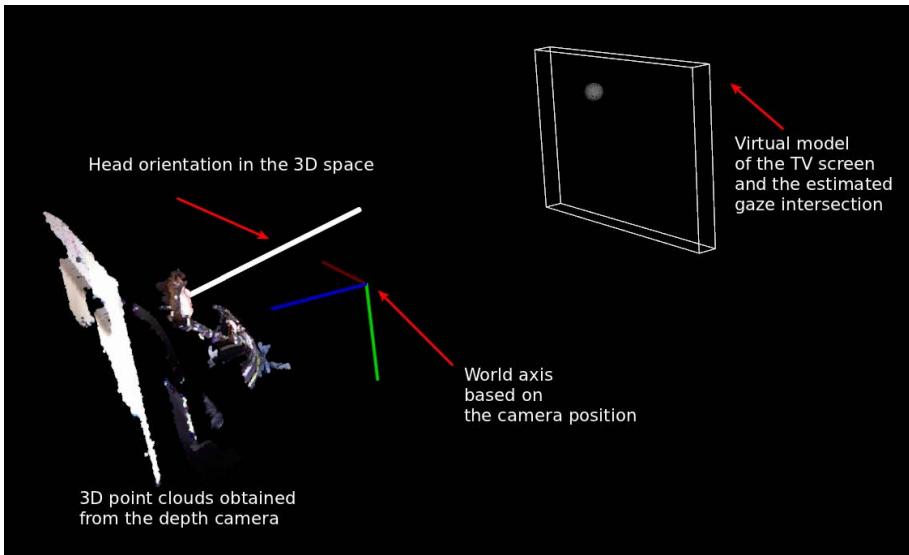


**Fig. 3.** Attention tracker workflow: 3D head pose and face recognition result go to the user profile and TV behaviour analysis which proceed to information low-level processing and fusion and forward it to the network manager module. The network manager takes all the messages (from the interest module, context tracking module and gesture module) and sends them to the player using the websockets protocol. The player enriches the messages with the video ID and time and forwards to the GAIN module that will aggregate the data.

#### 4.1 MetAttention: Image Metadata Based on a Visual Attention Mechanism

If it is possible to identify where a user approximately looks at, this information can be supplemented by bottom-up attention induced by the media. To investigate this kind of attention and couple it with the observation done on the user's behavior, we implemented a computational attention mechanism to analyze the bottom-up stimuli sent by the media. This algorithm is based on a bottom-up attention mechanism using a multi-scale rarity [27].

There are three main steps. First, we extract low-level colour and medium-level orientation features. Afterwards, a multi-scale rarity mechanism is applied. Finally, we fuse rarity maps into a single final saliency map. Contrary to RGB



**Fig. 4.** 3D rendering of our system. On the left, we can observe the 3D point cloud obtained with the depth camera. The head pose estimation algorithm is applied on this cloud, if a face is detected, we retrieve a vector of the head direction and compute an estimation of where the user is watching on the virtual screen.

color space, some alternative colour spaces (in our case YCbCr) give better uncorrelate colour information. Moreover, the nonlinear relations between their component are intended to mimic the nonlinear response of the eye. At this stage, the algorithm split in two pathways. The first one, mainly deals with colours (low-level features) while the second one with textures (medium-level features). While the first pathway directly uses the colour transformation and computes its rarity, the second pathway extracts orientation features maps by using a set of Gabor filters. These filters were chosen because they are similar to simple cells of the visual cortex in the brain [9]. For more information about the attention mechanism which was partly implemented, the reader can refer to [27]. The algorithm was implemented in C++ using OpenCV and multithreading, the performance in comparison of the Matlab implementation is 10x.

The output of the algorithm provides us with a map of saliency, using different steps of filtering and morphological operations it is then possible to segment the image into areas with high saliency values by using an adaptive thresholding. These areas will allow us to generate regions of bottom-up interest we can therefore correlate with measures of user's head direction obtained using the attention tracker.

Figure 5 shows the main processing steps of the algorithm:

1. The original image is converted in the YCbCr color space;
2. On each color channel a set of Gabor filters are applied to extract direction information and after that a multi scale rarity mechanism is applied;

3. The 6 rarity maps are fused together into a single map. This fusion is achieved in two main steps: an intra-channel fusion followed by an inter-channel one. The result is called a saliency map.

## 5 TV Web Player and Aggregator

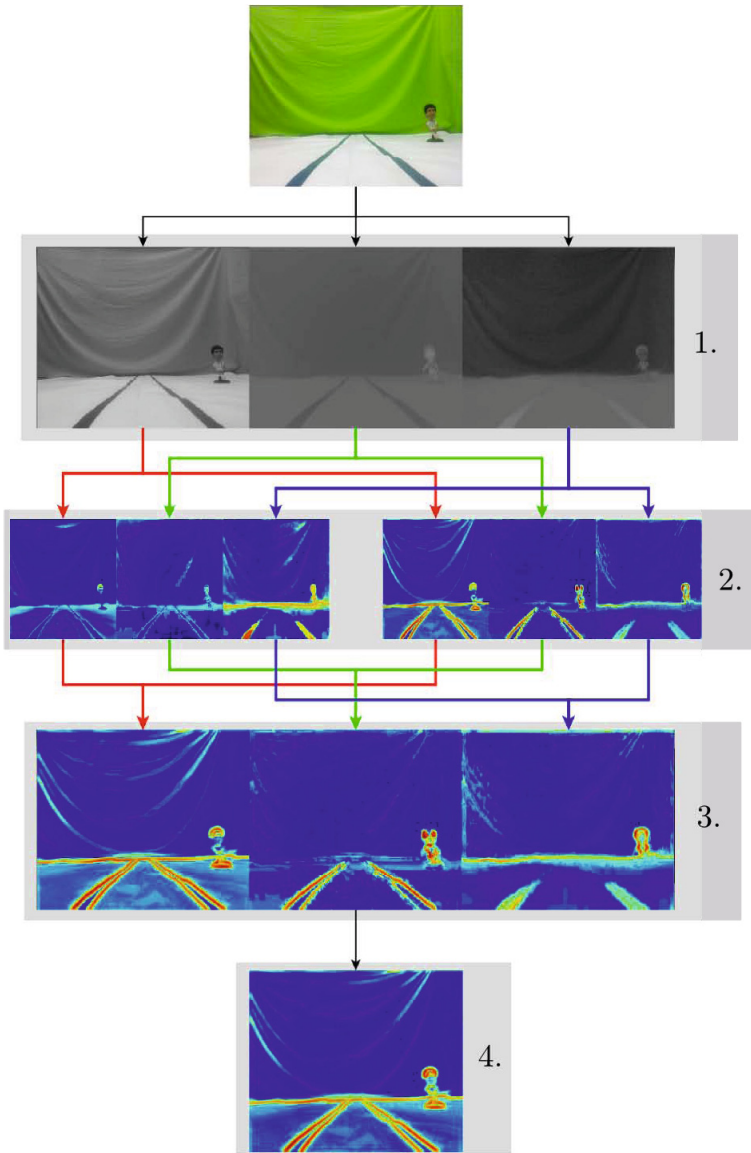
Figure 6 depicts the simplified workflow and communication of the modules at different levels. There are three levels: Web Browser, Server, Sensors. The application on the Sensors level was implemented in C++ and communicates to the server using Websocket protocol (sections 3 and 4). The server is implemented in Node.js and the application in the browser is implemented in HTML and JavaScript that communicates with the server using the Websocket protocol and REST API.

### 5.1 Player

The Player simulates the Smart TV environment using videos from YouTube (Figure 7). It is implemented as a web based application within a web browser. The interface provides the main screen with the video player, basic controls of the player and a semantic description of content based on analysis of subtitles. The viewer can interact with video using basic controls buttons or the user can read related content by clicking on the links to find more related information. Both Player and sensors for attention tracking and sensors for gesture control are connected using Websocket protocol. All detected gestures and interest clues detected by the attention tracker are sent to the synchronization service on the server. This information is propagated in nearly real-time to the Player. The player translates the incoming message into proper actions. Gestures control the player and attention information change is complemented with the video ID and video time which is displayed at the moment where the attention change occurs. All of these interactions, including actions from sensors, are sent to the GAIN component (section 5.2) using a REST interface.

### 5.2 GAIN - General Analytics INterceptor

GAIN ([www.inbeat.eu](http://www.inbeat.eu)) is a web application and service for capturing and pre-processing user interactions with semantically described content. GAIN outputs a set of instances in tabular form suitable for further processing with generic machine-learning algorithms. GAIN is implemented in Node.js and it is composed of three modules. First, a tracking module is responsible for capturing information. Afterwards, the storage module accumulates all data within a database. Finally, the aggregation module process the data and provides the outputs. GAIN provides RESTful API for collecting information and for aggregated outputs. GAIN will provide a timeline of the displayed video containing all synchronized actions which occurred during the viewer TV experience. The synchronized data contain both explicit (user actions like clicks, play, stop, etc.) and



**Fig. 5.** Diagram of our used model. First, from the input image, a change of color space is applied to improve the color differentiation (1.). Colour and orientation features are extracted. Then, for each feature, a multi-scale rarity mechanism is applied (2.). Finally, two fusions (intra- and inter-channel) (3.) are made from the rarity maps to provide the final saliency map (4.).



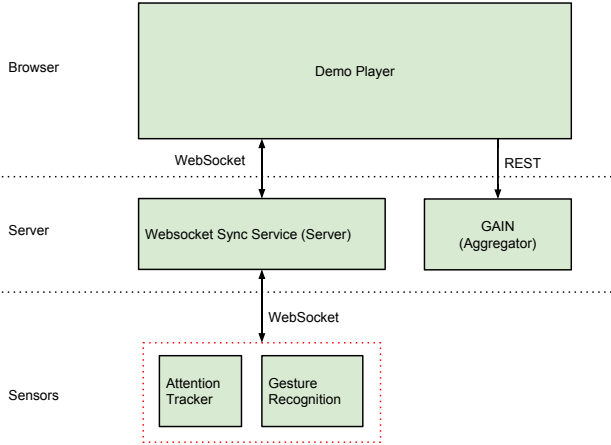


Fig. 6. Workflow

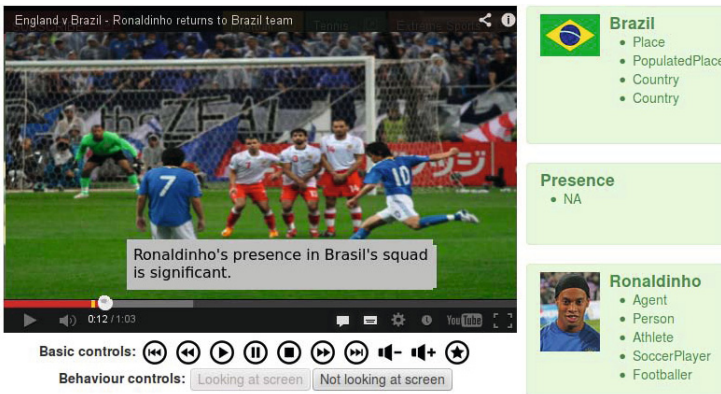


Fig. 7. An extended version of the YouTube player used for the SmartTV demo

implicit information (user behavior observation through the attention tracker). Examples of the output data can be seen in figures such as Figure 10.

## 6 Experimentation

To validate our approach of change detection in the viewer focus of attention, we designed an experiment leading to a moment of focus on a specific period of the broadcast media. Our goal was to determine whether it was possible for us to detect this moment through our system, but also to observe whether different behaviors stood out in conducting the experiment.

The experience consisted in watching 4 different videos in English and subtitled, through our web player (section 5). Each video has a different content and dealt with topics like politics or sports. While viewing these videos, additional content is provided to the user in real time on the web player (Figure 7 on the right). The subject were asked to be soccer fans and we asked them to watch the moments dedicated to this topic on the different videos which were displayed on the web player. To check that they really pay attention to soccer information, the users were asked to answer a questionnaire about relevant videos. Finally, to also cause changes in the individual attention we asked them to fulfil two different tasks while viewing the media:

1. The first task was to play a simple puzzle game on a tablet (IPad) and get the maximum possible score within the time limit.
2. The second task was to answer a series of questions on soccer dedicated to the broadcast videos. The stimulated interest was soccer and concerned only videos 2 and 4. Some questions had a higher difficulty push the user to exploit the extended content (access to Wikipedia which is available on the web player on the right side of Figure 7) to correctly answer the questionnaire.

## 6.1 Course of the Experiment

Our experiment took place in two stages. First, given the complexity of the operations to be performed, a tutorial simulating the experience is proposed to the subject. He could for 4 minutes try a simulation of the experience and learn how to handle the different elements that are provided (play/pause interactions, questionnaire and web player). At the end of this training phase, the actual experiment was performed with a time limit of 7 minutes. When the experiment was completed, the system stops and we collect the questionnaire and the game score.

The displayed media order is as follows:

1. The first video (2 min 18) is about a new US and South Korea joint defense plan against any provocation from North Korea and the help the US can provide to his ally. (url: [https://www.youtube.com/watch?feature=player\\_embedded&v=k4JstBd0sgk](https://www.youtube.com/watch?feature=player_embedded&v=k4JstBd0sgk)). This video has no information about the user simulated interest (soccer).
2. The second video (1 min 03) is about soccer, precisely about an England versus Brazil match and gives information about important player in the teams. (url: [https://www.youtube.com/watch?feature=player\\_detailpage&v=do5NcLT-t3s](https://www.youtube.com/watch?feature=player_detailpage&v=do5NcLT-t3s))
3. The third video (1 min 08) is about tourism and the 10 top attractions in Berlin. (url: [https://www.youtube.com/watch?feature=player\\_embedded&v=f9Uxzvekgio](https://www.youtube.com/watch?feature=player_embedded&v=f9Uxzvekgio))? Again, this video provides no information on the user simulated interest.
4. The fourth video (1 min 03) is about soccer and the transfer of a North Korean striker to South Korean club. (urel: [https://www.youtube.com/watch?feature=player\\_detailpage&v=v04ZM8HG4yg](https://www.youtube.com/watch?feature=player_detailpage&v=v04ZM8HG4yg))

**Interface KInterestTV (user interests: soccer)**

**\*Obligatoire**

**Participant Identification \***

**Who are the fans that support the Brazilian players during the training?**

Visual - look at the fence (2 points)

Brazilian president  
 English queen  
 Ordinary people holding Brazilian flag  
 Tom Mitchell

**When did Neymar win South American Footballer of the Year award?**

see Wikipedia (2 points)

2001  
 2008  
 2011  
 2012  
 2013

**What is the name of the Korean Football league?**

Kora League  
 K-League  
 Korean Soccer League  
 Korean Extra League  
 League Korean

**Where did Jong Tae Sae start his career?**

in Germany  
 in another club in South Korea  
 in Japan

**Number of points from the game \***

**Fig. 8.** Questions given to the user. The stimulated interest here is about soccer.

## 6.2 Interest Beats and Results

At the end of the experiment, 10 people have used our system and responded to the questionnaire. Based on the attention data obtained by our system, we generated a visualization of actions taken by the user that we call the "Interest beat". The expected behavior was the following and can be observed on the Figure 10:

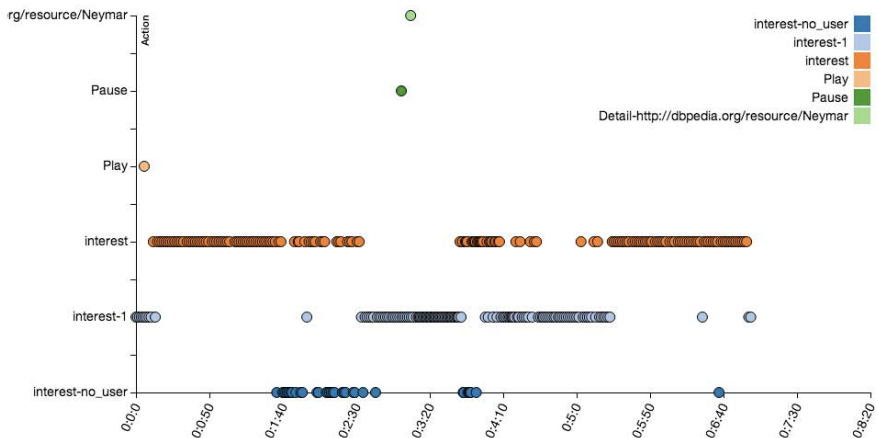
1. While the videos not related to the questionnaire are playing (1st and 3rd), the user will mainly focus on the game on the second screen (and therefore will not look at the main screen)
2. While the videos related to the questionnaire are playing, the viewer looks to the main screen and therefore stops playing the game on the second screen. He can also stop or jump in this video to go back to a topic related to the questions of their questionnaire

The interest beat is a graph presenting the output of the GAIN module [21] and consists in time synchronized events happening all over the content (X axis)



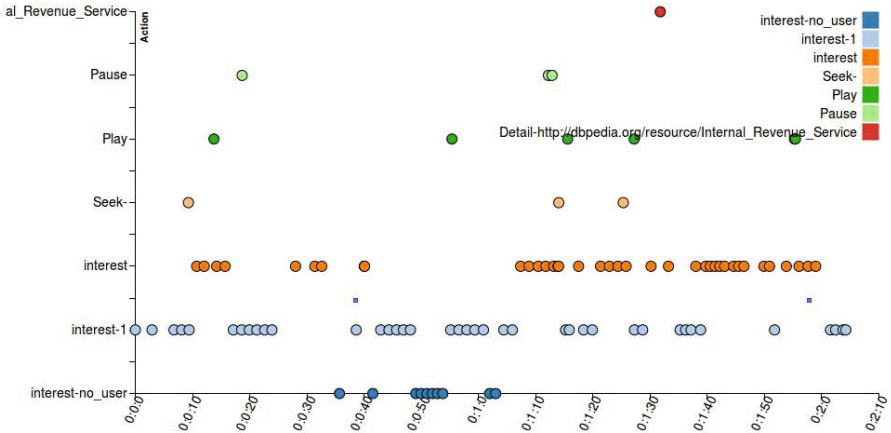
**Fig. 9.** Experimental setup: videos are displayed on a computer screen while the user needs to play a game and answer questions simultaneously. A 3D sensor on top of the screen captures the head movement of the user.

## Timeline



**Fig. 10.** Interest beat. The timeline comes from each event while the experiment. It represents the viewer behaviour over the displayed content (first video from 00:00:00 to 00:02:18, second video until 00:03:21 and the third video until the end 00:04:24). On the Y axis: clicks on links, pause, play, interest-0 (not looking at the main screen), interest -1 (looking at the main screen) and interest-no-user (user not found anymore).





**Fig. 12.** Interest beat with a user alternatively looking at the main screen and at the game

on link, etc.) were performed, the other users have a significant activity towards the main screen. Nevertheless, it is not easy to assess the tracking performance as the users sometimes have different behavior. Sometimes the viewer forgot to answer the questions, so he had to go back in the video to do it at the end like in Figure 11.

In other users, they are very consistent with the task (playing or looking to the screen like in Figure 10 while other alternate much more the gaze between the main screen and the game like in Figure 12. On the nine viewers where the tracking worked, the results are consistent with the scenario and encouraging.

## 7 Conclusion

In this paper, we presented the whole architecture of our implicit behavior analysis system based on a 3D head tracker and also the explicit gesture recognition system. We also describe the additional information which can be provided by the extraction of low-level bottom-up features from the media. The results show that it is possible to extract implicit information in an efficient and consistent way on where and when people look at their TV. Modules which provide a web player and a data aggregator are used to synchronize all the behavioral analysis of the viewer. This work is designed to further feed a personalization framework capable of processing behavioral data to dynamically enhance the profile of a user. This profile change needs further machine learning algorithms which take the synchronized data of the proposed system as an input and process it in order to obtain the user interest for the different media fragments and media links which were shown to the user.

**Acknowledgments.** This work is supported by the Integrated Project LinkedTV ([www.linkedtv.eu](http://www.linkedtv.eu)) funded by the European Commission through the 7th Framework Programme (FP7-287911).

## References

1. Linkedtv project, <http://www.linkedtv.eu>
2. Node js, <http://nodejs.org/>
3. Abe, K., Makikawa, M.: Spatial setting of visual attention and its appearance in head-movement. *IFMBE Proceedings* 25/4, 1063–1066 (2010), [http://dx.doi.org/10.1007/978-3-642-03882-2\\_283](http://dx.doi.org/10.1007/978-3-642-03882-2_283)
4. Aldoma, A.: 3D face detection and pose estimation in pcl (September 2012)
5. Bailly, G., Vo, D.B., Lecolinet, E., Guiard, Y.: Gesture-aware remote controls: Guidelines and interaction technique. In: *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011*, pp. 263–270. ACM, New York (2011), <http://doi.acm.org/10.1145/2070481.2070530>
6. Bettens, F., Todoroff, T.: Real-time dtw-based gesture recognition external object for max/msp and puredata. In: *Proc. SMC 2009*, pp. 30–35 (2009)
7. Bevilacqua, F., Guédy, F., Schnell, N., Fléty, E., Leroy, N.: Wireless sensor interface and gesture-follower for music pedagogy. In: *Proceedings of the 7th International Conference on New Interfaces for Musical Expression, NIME 2007*, pp. 124–129. ACM, New York (2007), <http://doi.acm.org/10.1145/1279740.1279762>
8. Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001), <http://dx.doi.org/10.1023/A%3A1010933404324>
9. Daugman, J.G.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A* 2(7), 1160–1169 (1985), <http://josaa.osa.org/abstract.cfm?URI=josaa-2-7-1160>
10. Fanelli, G., Gall, J., Van Gool, L.: Real time head pose estimation with random regression forests. *Cvpr 2011*, 617–624 (2011), <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5995458>
11. Fanelli, G., Gall, J., Van Gool, L.: Real time 3d head pose estimation: Recent achievements and future challenges. In: *2012 5th International Symposium on Communications Control and Signal Processing (ISCCSP)*, pp. 1–4 (2012)
12. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Gool, L.: Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision* 101(3), 437–458 (2012), <http://link.springer.com/10.1007/s11263-012-0549-0>
13. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Gool, L.: Random forests for real time 3d face analysis. *International Journal of Computer Vision* 101, 437–458 (2013), <http://dx.doi.org/10.1007/s11263-012-0549-0>
14. Fanelli, G., Weise, T., Gall, J., Van Gool, L.: Real time head pose estimation from consumer depth cameras. In: Mester, R., Felsberg, M. (eds.) *DAGM 2011. LNCS*, vol. 6835, pp. 101–110. Springer, Heidelberg (2011)
15. Venturini, F., Marshall, C., Di Alberto, E.: Hearts, minds and wallets winning the battle for consumer trust accenture video-over-internet consumer survey (2012)
16. Frisson, C., Keyaerts, G., Grisard, F., Dupont, S., Ravet, T., Zajga, F., Colmenares-Guerra, L., Todoroff, T., Dutoit, T.: Mashtacycle: On-stage improvised audio collage by contentbased similarity and gesture recognition. In: *5th International Conference on Intelligent Technologies for Interactive Entertainment, INTETAIN (2013)*

17. Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruyter, J., Knoll, A.: Social behavior recognition using body posture and head pose for human-robot interaction. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2128–2133 (October 2012), <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6385460>
18. Kistler, F.: Fubi- full body interaction framework (2011), <http://www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/fubi/>
19. Kistler, F., Endrass, B., Damian, I., Dang, C., Andr, E.: Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces*, 1–9, <http://dx.doi.org/10.1007/s12193-011-0087-z>, doi:10.1007/s12193-011-0087-z
20. Kistler, F., Sollfrank, D., Bee, N., André, E.: Full body gestures enhancing a game book for interactive story telling. In: Si, M., Thue, D., André, E., Lester, J.C., Tanenbaum, J., Zammitto, V. (eds.) ICIDS 2011. LNCS, vol. 7069, pp. 207–218. Springer, Heidelberg (2011)
21. Kuchař, J., Kliegr, T.: Gain: Web service for user tracking and preference learning - a smart tv use case. In: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys 2013, pp. 467–468. ACM, New York (2013), <http://doi.acm.org/10.1145/2507157.2508217>
22. Leroy, J., Rocca, F., Mancas, M., Gosselin, B.: 3D head pose estimation for tv setups. In: Mancas, M., d' Alessandro, N., Siebert, X., Gosselin, B., Valderrama, C., Dutoit, T. (eds.) Intetain. LNICST, vol. 124, pp. 55–64. Springer, Heidelberg (2013)
23. Leroy, J., Rocca, F., Mancas, M., Gosselin, B.: Second screen interaction: An approach to infer tv watcher's interest using 3d head pose estimation. In: Proceedings of the 22nd International Conference on World Wide Web Companion, WWW 2013 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 465–468 (2013)
24. Microsoft: Kinect sensor, <http://www.xbox.com/kinect>
25. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(4), 607–626 (2009), <http://www.ncbi.nlm.nih.gov/pubmed/19229078>
26. PrimeSense: Capri sensor, <http://www.primesense.com/news/primesense-unveils-capri>
27. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., Dutoit, T.: Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication* 28(6), 642–658 (2013), <http://www.sciencedirect.com/science/article/pii/S0923596513000489>
28. Vatavu, R.: A comparative study of user-defined handheld vs. freehand gestures for home entertainment environments. *Journal of Ambient Intelligence and Smart Environments*
29. Vatavu, R.D.: User-defined gestures for free-hand tv control. In: Proceedings of the 10th European Conference on Interactive Tv and Video, EuroITV 2012, pp. 45–48. ACM, New York (2012), <http://doi.acm.org/10.1145/2325616.2325626>
30. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27(12), 1743–1759 (2009), <http://www.sciencedirect.com/science/article/pii/S0262885608002485>
31. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2009, pp. 1083–1092. ACM, New York (2009), <http://doi.acm.org/10.1145/1518701.1518866>



# Development of an Ecosystem for Ambient Assisted Living

João Rosas, Luis M. Camarinha-Matos, Gonçalo Carvalho, Ana Inês Oliveira,  
and Filipa Ferrada

Universidade Nova de Lisboa, Faculty of Science and Technology, Campus de Caparica,  
2829-516 Monte Caparica, Portugal  
{jrosas,cam,aio,faf}@uninova.pt,  
g.carvalho@campus.fct.unl.pt

**Abstract.** Society is facing big demographic changes. In 2050, it is expected that the number of elders will reach 1500 million (about 16% of the world population). As people age, they become more dependent on assistance services. Care and assistance organizations start failing, as the number of people who need help increases beyond their ability to comply. The creation of an ecosystem for Ambient Assisted Living, facilitating partnerships creation between service providers, is proposed as a strategy to improve care provision and leverage its capacity. The specification of the ecosystem is based on canonical models and verified through simulation.

**Keywords:** Ambient Assisted Living, Ecosystem, Collaborative Networks, Information and Communication Technologies, Simulation.

## 1 Introduction

The society where we live is facing a big demographic change. People live longer, therefore life expectancy is increasing. In 2000, there were already 420 million people with more than 65 years old (which corresponded to about 7% of the world population). In 2050, it is expected that this number reaches 1500 million (about 16% of the world population) [1, 2]. The number of elder people who needs care and assistance is also increasing, surpassing the number of youngsters who contributes with taxes.

This context brings new challenges to the traditional health care systems, as social security systems are becoming unable to afford the cost of providing assistance to this growing number of people. Therefore, there is an increasing necessity to search for new solutions that will allow people to live in the best possible way, in the last stages of their life. These systems would allow people to extend their life in their favorite environments, favoring confidence, autonomy, mobility and welfare.

Information and communication technologies (ICT) offer new opportunities for the provision of improved care and assistance services. Ambient Assisted Living (AAL) is a concept focused on the use of technology as a way to improve the independence and welfare of aged or disabled people, at their homes.

With this research work, we aim at contributing to provide an answer to this need. Our goal consists of developing an Ecosystem for AAL. Our strategy is based on

Collaborative Networks [3]. For such, our effort was focused on the instantiation of the conceptual architecture proposed in the AAL4ALL project [4]. We relied on the utilization of canonical models for the development of our system, allowing the formulation of more simple, yet useful, specifications. We tested the AAL ecosystem as a distributed simulation system. In this regards, we take simulation as a design paradigm [5]. Within this paradigm, we can predict how a system is likely to behave in the future. We can also test the system when it is subjected to specific situations, e.g. the failure of services provision, too many assistance services requests or infrastructure shutdown, and assess if it works adequately. Given that many of these failures can only be experienced after the system is put in operation, using simulation, we can anticipate them and modify the specification of the system, prior to the physical development.

This research work was performed during the eNTERFACE'13 summer workshop [6]. Given that the timeframe of this event is a single month, many of the features we started to develop during the workshop are continued in the context of the AAL4ALL project.

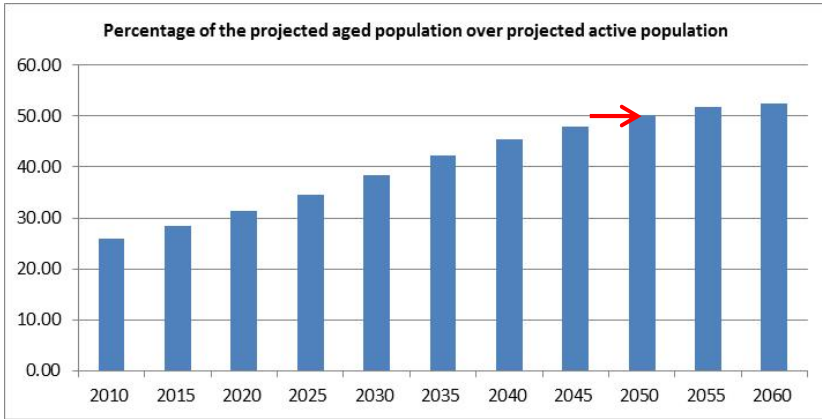
We start this work by firstly presenting an overview of the AAL concept and its most important aspects. In section 2, we present a literature review, mostly concerned with introducing and analyzing current aspects about assistance services, users, service providers, sensors and actuators, and supporting infra-structure. Section 3 is devoted to establishing the requirements, specification development and validation of the proposed AAL ecosystem. Finally, section 4 provides a synthesis of the work, achieved results and proposes the next steps for future work.

## **2 Ambient Assisted Living**

The area of AAL can be addressed from several perspectives, as it comprises technological, strategic, economic, social, moral and regulatory aspects. During the literature review, we emphasize some of the mentioned aspects, aiming to provide an adequately general overview of AAL. Through this chapter, the current developments in the field of AAL, namely in terms AAL products, services, and available platforms are illustrated. Relevant international projects and roadmaps are also mentioned. In this regard, this work can also be seen as a contribution towards the effective development of solutions that may help to better deal with this demographic trend.

### **2.1 The Importance of AAL**

As mentioned before, there has been a severe demographic change, faced by most developed countries, which leads to a rapid increase of the percentage of aged population. As illustrated in Fig. 1, the projected percentage of people above 65 years old over the active population reaches 50% by 2050. This trend of having fewer youngsters, who have to support the growing elderly population, requiring increased assistance, raises costs and leads to the rupture of the social security and social care systems.

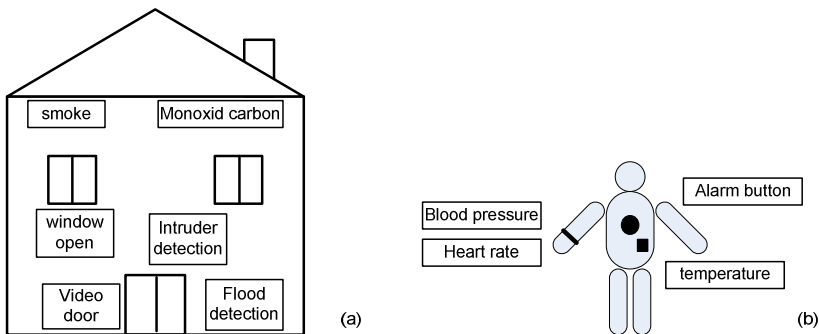


**Fig. 1.** Projected number of persons aged 65 and over expressed as a percentage of the projected number of persons aged between 15 and 64 (statistical data from [7])

In order to maintain the care provision capacity at an affordable level in the future, there is an urgent need to find effective and affordable solutions to provide care and assistance to elderly.

## 2.2 The AAL Concept and Technology

Ambient Assisted Living is a concept in which technology is used as a way to improve the welfare and independence of elder or disabled people living alone at their homes. Typically, a variety of sensors and actuators installed at their homes or in their clothes are used to remotely monitor their wellbeing conditions. Fig. 2 illustrates a home and a user, which contains a number of sensors for remotely monitoring users' welfare. These devices operate supported by an infra-structure, usually of wireless



**Fig. 2.** AAL environment: (a) environmental sensors; (b) physiological sensors

type, which provides adequate connectivity. It is on top of these devices that AAL services operate.

Typical services in AAL include home environment services, like home safety and security, temperature monitoring, gas detection, smoke detection, intruder alert, fall detection. It can include welfare monitoring for people that are not ill, like monitoring heart rate, blood pressure, and body temperature. It can also include health support for ill people, like behavior monitoring (for people with dementia) and chronic disease management. Furthermore, it may also include occupation and recreation services, which support the involvement of leisure services and the continuation of professional activities.

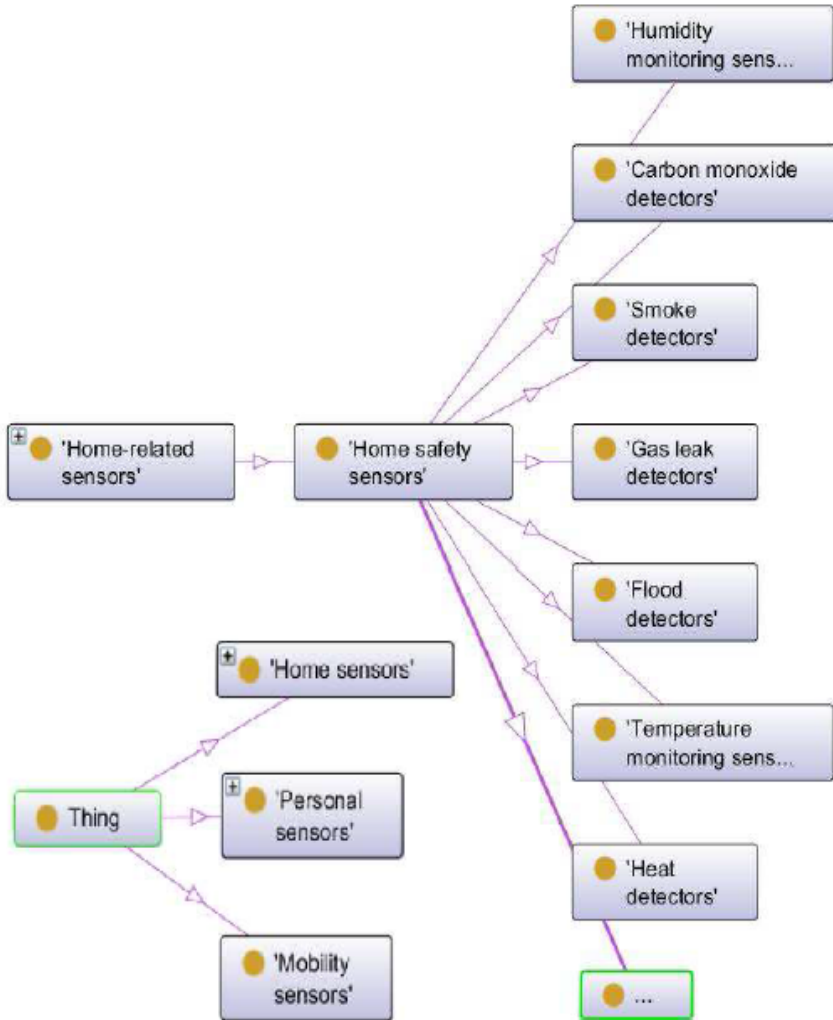
### 2.3 AAL Sensors and Actuators

AAL sensors are used to monitor the health and welfare situation of the users. If these sensors are combined with communication devices, usually wireless, we can monitor the users remotely. There are many types of AAL sensors, namely, blood pressure, cardiac, body temperature, pedometers, urinary salinometer, and so on. According to their characteristics, more specifically, the type of performed observations, they can be split into several categories. For instance, sensors can be classified as “home safety sensors”, which are used to monitor and protect from internal threats and ensure that users' homes are in safe conditions, regarding aspects such as temperature, smokes, floods, or gas leakages. Another sensor category, “home security sensors”, help protect users from external threats, e.g. for monitoring intrusion, detection of presence, or doors/windows opening. As illustrated in Fig. 3, home safety sensors hold a large number of sensor sub-classes, each one devoted to each aspect of users' wellness, like monitoring home temperature, smoke detection, gas leakage, and carbon monoxide. These varieties can be organized in taxonomies.

As opposite to home sensors, AAL sensors used to monitor users welfare or health can be classified as personal, which are used to monitor users welfare conditions.

Such great variety of types and technologies in sensors raises concerns related to integration and interoperability, because the available electronic devices on the market do not provide standardized interfaces for integration. In a further section, we mention a number of research projects, which deal with integration and interoperability in AAL. There are also currently focused concerns related to the privacy of the data collected to monitor the elderly [9].

AAL actuators can perform interventions on the elders or on their homes, in order to adjust elderly welfare. If these actuators are equipped with communication devices, usually also wireless, interventions can be performed remotely. Fig. 4 illustrates a partial taxonomy of AAL actuators.



**Fig. 3.** Sensor taxonomy illustrating several types of AAL sensors (Note: the “Thing” element is automatically included by the ontology editor [8])

According to each purpose, AAL actuators can be classified into broad categories in a similar way as the AAL sensors. There are, nevertheless, huge concerns in the utilization of AAL actuators because they are devices which allow remote changing of the state of the elder or his home, and this poses many security issues, such as: adequate performance, operation by non-authorized personal, etc., that may lead to harmful situations.

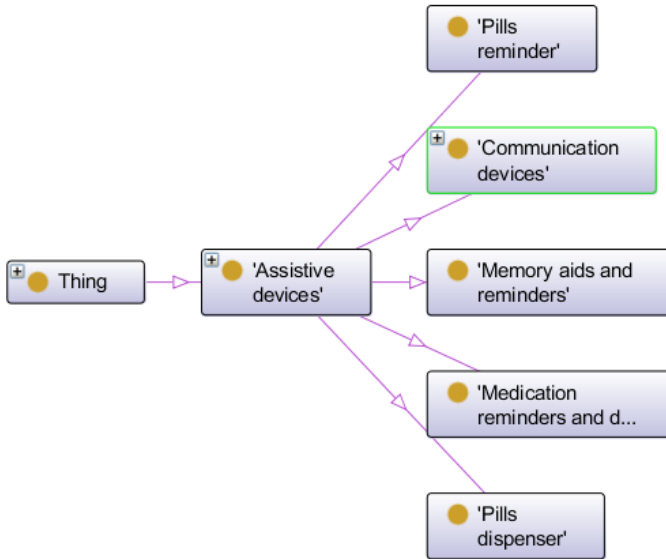


Fig. 4. AAL actuators

## 2.4 AAL Services

AAL services intend to enable people with specific needs to live an independent and save life. To this end, the services have to cope with a number of user and environmental challenges [10], namely, low or declining capabilities, distinct elderly needs accordingly to health conditions, limited (human and financial) resources, low tolerance of technical problems, the desire to feel in control of their lives, avoid stigmatization and keeping privacy.

As mentioned in [11], past research and developments in elderly care services, as well as current market offers, are characterized by some fragmentation. The focus has been predominately put on the development of isolated services, e.g. monitoring of some health related parameters, fall detection, agenda reminder, alarm button, etc., each one typically provided by a single organization, and often showing an excessive techno-centric flavor. Contrasting to this situation, the concept of Care and Assistance Service (CAS) is proposed in [11], which refers to a category of services, either of a medical or social nature, aiming at helping senior citizens in their daily lives, compensating for the reduction of physical and/ or mental capabilities that comes with the ageing process. As mentioned in [11], the execution of a care and assistance service may involve a number of software services and human intervention (manual tasks). The actual structure of such service also depends on the interaction between the provider and the end-user, and may ultimately (and dynamically) vary according to the flow of that interaction.

Similarly to sensors, AAL services exist in multiple categories. In AAL4ALL, four categories of services, named as services for four specific life-setting, were considered, as identified in Fig. 5.

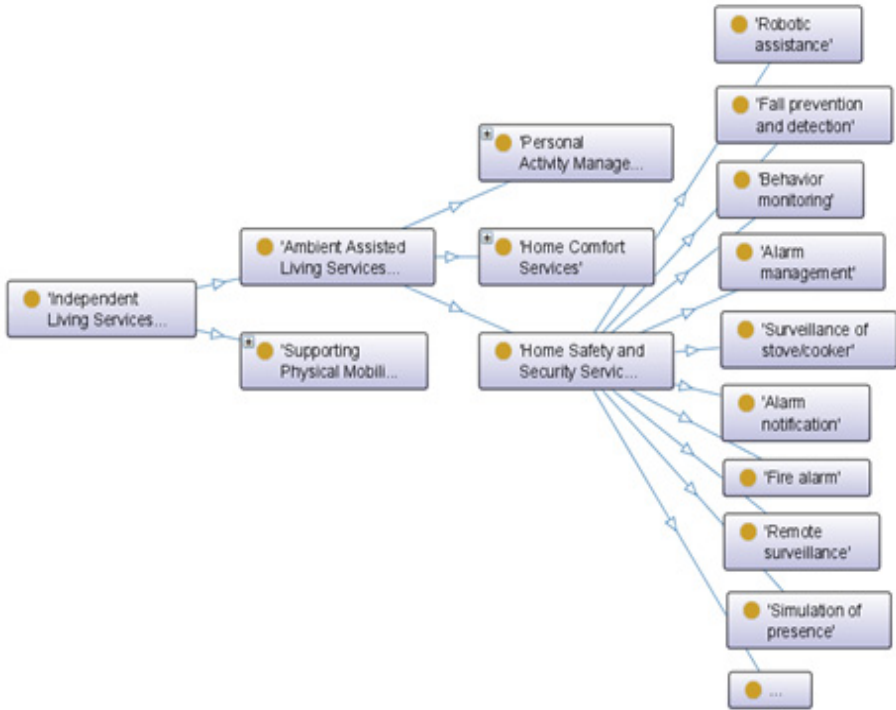


Fig. 5. Illustration of *home safety and security* services category, proposed in the AAL4ALL

## 2.5 AAL Service Providers

Providing AAL services involves the participation of several types of entities, ranging from health care providers, to day care centers, to hospitals and fire departments. ICT providers are responsible for supplying of necessary technological infrastructures, which allows AAL services providing.

An important aspect, which is in fact addressed during this research work, is that in the AAL context, given the heterogeneity of these stakeholders, a single and isolated provider is not enough to satisfy every concrete need of AAL users. As described in Section 3 of this article, the proposed approach is to develop an AAL ecosystem, which allows providers to collaborate and establish partnerships. In such way, they can together be able to provide the adequate AAL services to each user.

These services require adequate ICT infrastructures for a proper service provision. Such infrastructure involves data models, hardware and software components, processes and people. Information flow includes observations from sensors, user profiles, services providers' interactions and treatments. Additionally, a typical AAL environment includes heterogeneous components of distinct technology, which need to interoperate. A useful approach for such interaction between these components is to rely on a service-oriented approach, such as the Enterprise Service Bus [12,13], in

which the elements of heterogeneous nature are able to interact using a standardized protocol.

## 2.6 Current Research in AAL

**Research Projects.** As mentioned in [14], a considerable number of research projects are focused on developing systems that monitor the health or welfare conditions of the elder and their activities, looking for disturbances requiring assistance by the service providers. The most common architecture is based on a set of sensors, more or less extended, dispersed throughout the housing and allowing the control of a number of activities (for example, opening and closing of doors, stove utilization or the use of refrigerator water consumption / gas / electricity, etc.). By merging all the information gathered by different sensors, it is possible to determine the current activity of the person who is being controlled and identify deviations from his/her daily routines. These deviations can be used to detect emergency situations or changes in the condition of the person.

Internationally, there are several projects that have addressed these aspects. For instance, the projects: UbiSense [15], ROSETTA [16] and Dreaming [17] focused on the development of systems for elderly health and welfare conditions in order to keep their independence at their homes. The projects ITALH [18] and OASIS [19] addressed the interoperability issues between distinct technologies (zigBee, Bluetooth, GPRS, etc.). The i2Home project [20] was devoted to the development of new devices for AAL based on existing industrial standards. The AWARE [21] developed a social network for promoting the social inclusion of the elderly, aging workforce and contribution to society.

At the European level, there were several Roadmap projects in the area of AAL. For instance, the AALIANCE project [22] was devoted to the creation of innovative AAL devices and sensors. The ePAL project [23] was about promoting balanced and active life for retired people or in the process of retirement in Europe. The SENIOR project [24] was concerned with performing a systematic assessment of the social, ethical and privacy concerns in aging and ICT. Finally, BRAID project [25] addressed the research and Technological Development for active ageing, through the consolidation of the results from other existing roadmaps.

**Existing AAL Products.** The market of AAL products is very fragmented, as there are still not fully consolidated AAL solutions that meet the needs of all countries worldwide. Although they are still significantly expensive, it is expected that the products in the AAL area will decrease about 50% of the costs associated with health care services for senior people, envisaging, for example, that the U.S. market of AAL is in 20 million euros per year with a rising trend [14].

There are many products already available in the market. For instance, Equivital [26] is a system composed of several monitoring sensors connected to a wearable wireless module. Sensium [27] is a platform which continuously monitors a user's body using non-intrusive wireless sensors. The Hallo Monitoring [28] is a fall



detector with automatic alarm sending service. Grand Care [29] is a system described used to monitor the daily activity and welfare of a person in a non-intrusive way. HomMed [30] is a monitoring unit placed at a person's home, which collects and transmits data on the person's health conditions to a service center that processes and presents this information to health care providers.

**Critics to Existing Approaches.** As we mentioned in [3], many of the AAL initiatives are characterized by being too techno-centric, without properly addressing social and strategic aspects. On the other hand, AAL services are too fragmented and provided by different service providers. These providers are also characterized by a deep heterogeneity. However, no single AAL provider can adequately fulfill the elderly needs. Given these factors, and as already mentioned, the best strategy for the provision of AAL would be through collaborative approaches.

As such, a trend identified in a recent work [3], shows the need to move from a scenario characterized by fragmented services, typically provided by single service providers, and often showing an excessive techno-centric flavor, to more integrated care services. In contrast, there is now a perception that is fundamental, for the success of future AAL support systems, to seek synergies between the areas of ICT, Ageing and Collaborative Networks. Integrated AAL services would then be provided by multiple stakeholders, through partnerships.

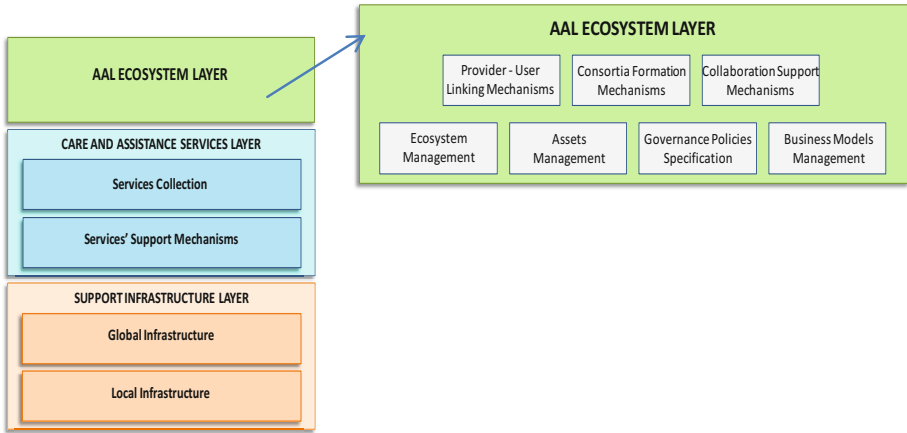
### 3 AAL Ecosystem Development

Considering the several aspects revised in previous chapters, we describe the specification and development of the AAL ecosystem support platform in this chapter. As suggested from the international trends, described on the BRAID project [25], the approach for creating the ecosystem is based on collaborative networks.

#### 3.1 The AAL4ALL Conceptual Architecture

In order to adequately handle the mentioned issues, the adopted approach is to instantiate the AAL4ALL architecture illustrated in Fig. 6. We adopted the vision shared in AAL4ALL project of following a more socio-technological approach. As such, we are more focused on instantiating the top layer of the architecture, which is used to assist us in the specification and implementation of the AAL ecosystem that is proposed in this work.

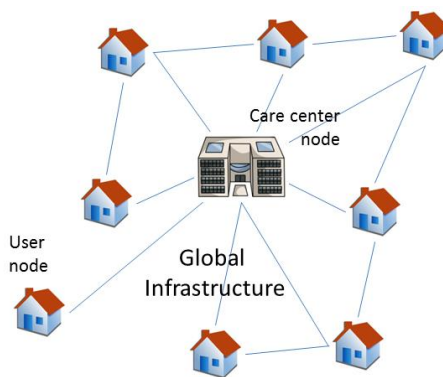
This architecture illustrates the adopted perspective of focusing on the socio-business aspects at the top layer, by providing functionality for partnership creation. The architecture is composed of the AAL ecosystem layer, the services layer, and the infrastructure layer. The ecosystem layer lies at the top of the architecture. In this layer, it is important to consider the management and governance principles of AAL ecosystem. It is based on a collaborative network philosophy, facilitating an effective collaboration between the stakeholders participating in AAL service provision.



**Fig. 6.** AAL4ALL Conceptual Architecture, inspired and adapted from [4]

For the services layer, given the large scope of the area and the complexity of AAL, which is of a multi-disciplinary nature, it becomes convenient to consider complementary perspectives of analysis. In the AAL4ALL project, four different life settings in an elder's life were considered, namely: Independent Living, Health and Care, Occupation in Life, and Recreation in Life. These perspectives help adequately organize the collections of assistance services in AAL services taxonomies as illustrated in the previous chapter. Such collections allow fulfilling the envisaged necessities corresponding to the elderly conditions. Additionally, as services are of distinct characteristics, e.g. ambient monitoring services versus health monitoring services, they might be provided by distinct service providers. As such, the service collections are organized according to the mentioned four life-settings.

The infrastructure layer plays the role of a facilitator (provides support) for the development and delivery of care and assistance services. Such infrastructure should provide, among other functionalities, channels and mechanisms for safe



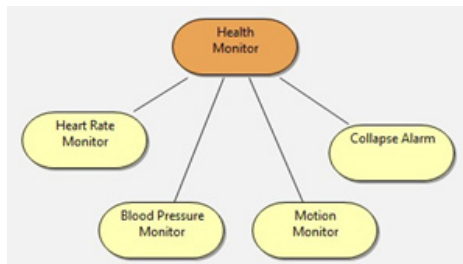
**Fig. 7.** Representation of a global infrastructure composed by several local infrastructures

communications and information sharing and exchange among the members of a given AAL ecosystem. It has two sub-layers, the local infrastructure corresponding to the support infrastructure located in a specific "location", e.g. users' home, care center, health care center, human-centered environment (intelligent cloth, mobile gadgets, etc.); and the global infrastructure, supporting the network of "spaces" (or local environments) "inhabited" by the various stakeholders. The global infrastructure illustrated in Fig. 7 supports the interaction between the entities/nodes engaged in care provision and the assisted people. It supports multi-node services, distributed processes, software services invocation and composition.

### 3.2 Requirements Identification

According to what was mentioned before, the adopted approach for the AAL ecosystem development is based on canonic models specification. This type of model allows the specification of a system with considerable complexity using simple, yet useful structures. This allows developing the system within a limited timeframe, without undermining the identified functional requirements.

**AAL Services.** Pursuing an approach based on canonical models, an AAL service can be seen as a set of actions designed to provide care to a user. A service can be identified by a name and a pre-established functionality. A service can be also composed for other services in a hierarchical composition structure, as shown in Fig. 8.



**Fig. 8.** Hierarchical services definition

This complies with the notion of service established in the previous chapter, and as illustrated in Fig. 5. The combination of services results in "tailored / customized packages". It is a way to meet particular needs of certain users. The functional requirements for the management of the AAL services are described in Table 1.

**Table 1.** Functional requirements of AAL services management

FR	Description
FR_s1	Creation of services which can be adequately characterized by ID, name and type
FR_s2	Search, update and deletion of existing services
FR_s3	Creation of services as a composition of other services
FR_s4	Contract of services by users
FR_s5	Association of a running service to a user or users' home

**AAL Users.** In canonical terms, a user can be characterized by a name, an address and a set of pre-established AAL characteristics (e.g. in an ontology). The functional requirements for the management of users are listed in Table 2.

**Table 2.** Functional requirements for the AAL users management

FR	Description
FR_u1	Creation of new users, which can be adequately characterized by ID, name and a set of pre-established attributes for its AAL characterization
FR_u2	Search, update and deletion of existing users
FR_u3	Service contracts establishment with users
FR_u4	Track of AAL relevant events from a users' services
FR_u5	Billing of contracted services

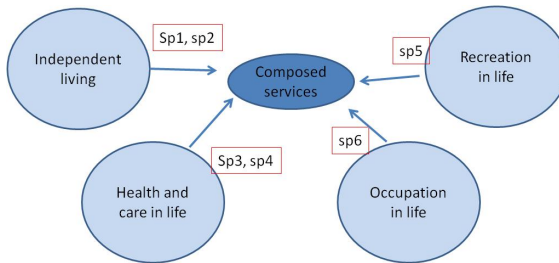
**AAL Service Providers.** Providers can be characterized by name, location, and type. (Infrastructures providers, and AAL service deliver, care centers, surveillance and security providers, etc.). The functional requirements of providers' management are identified in Table 3.

**Table 3.** Functional requirements for AAL providers management

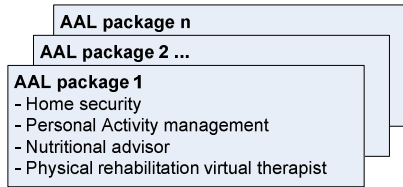
FR	Description
FR_p1	Creation of new providers, which can be adequately characterized by ID, name and a set of pre-established attributes for its AAL characterization
FR_p2	Search, update and deletion of existing providers
FR_p3	Creation of providers as composition of other providers
FR_p4	Association of providers and the services they deliver
FR_p5	Participation of providers in services contracts with users

**AAL Ecosystem.** Given that an elder, or a group of elderly, may have got specific necessities, they require distinct types of assistance services. In these cases, it is necessary to provide customized AAL services. However, a service provider operating alone is not usually able to provide the necessary services to these elderly. A solution would be a costly investment to obtain the capacity to provide the necessary services. As a better approach, when tailored care provision is needed, service providers can organize themselves in partnerships, creating composed services that suit these needs. For instance, as illustrated in Fig. 9, *independent living, health, occupation and recreation* service providers can join in order to combine their services.

Inside such partnerships, it is now possible to create customized services, or service packages, as illustrated in Fig. 10, which are able to suit the particular necessities of a number of users.

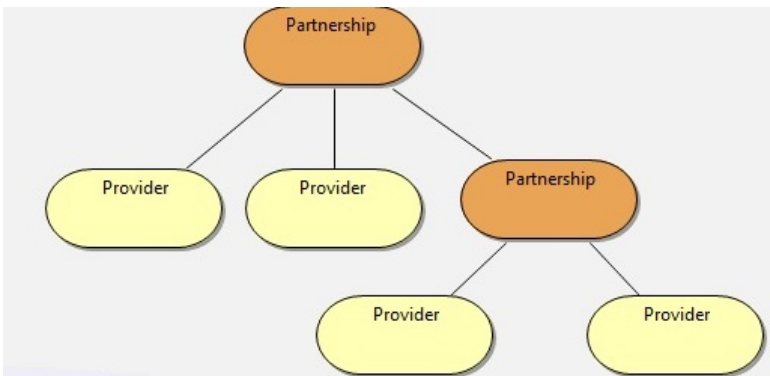


**Fig. 9.** Creation of composed services from providers of distinct types



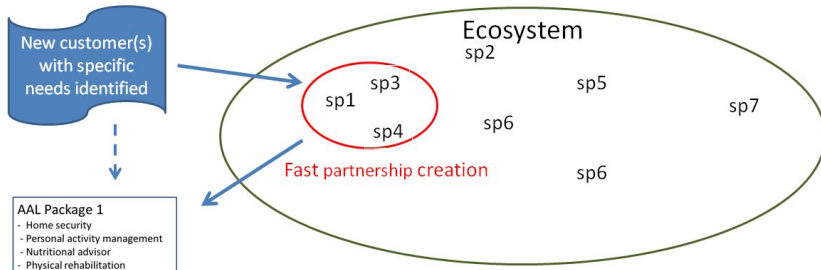
**Fig. 10.** Customized packages of AAL services

Following the approach of developing canonical models, we can assume that a partnership can also be seen as a single service provider. Therefore, a provider can be composed by a single entity or by several entities. Some entities participating in a partnership may, per se, be already a partnership. This leads to a hierarchical definition of service provider, as illustrated in Fig. 11.



**Fig. 11.** Single versus partnership service providers

The AAL ecosystem, illustrated in Fig. 12, is composed of several service providers, each one delivering their own services. It is a virtual space in which members (service providers) agree to create partnerships as soon as good opportunities are found. It facilitates fast, dynamic, on-the-fly partnership formation. As members of the ecosystem, partners remain prepared to engage in partnerships [31]. When a market opportunity for a new customized service appears, they engage in fast partnership creation. The partnership remains while the service package is being delivered.



**Fig. 12.** Illustration of an AAL ecosystem

Considering these aspects, the requirements for the AAL ecosystem are described in Table 4.

**Table 4.** Functional requirements of the AAL ecosystem management

FR	Description
FR_e1	Registering of users/elders
FR_e2	Registering of providers
FR_e3	Partnerships creation
FR_e4	Partnership life-cycle management
FR_e5	Business model (tailored packages, responsibilities, profit sharing approach, etc.)

### 3.3 Ecosystem Specification

The specification of the AAL Ecosystem requires the preliminary definition of a number of concepts, namely, the users, their homes, the Care Providers, and the AAL Services. These definitions will then be used in the specification of the ecosystem functionality. During the definitions set below, we use lower case for specifying single elements and upper case for sets.

**Definition 1 (AAL User)** - Can be an elder or a person that lives alone or with barely any assistance that wants to continue living in his own home. This user subscribes to one or more services that compensate his/her limitations, aiming to improve welfare and safety. A user can be abstractly defined as a tuple  $u = (ID_{user}, Name, V_{Attr})$ . The set  $V_{Attr}$  represents the attributes that adequately characterize a user. From now on, let us consider the existence of the set of  $U = \{u_1, u_2, \dots, u_n\}$ .

**Definition 2 (User’s Home)** - The environment where the user lives is also an important part of the AAL Ecosystem. If the user has limitations, it is important to monitor his/her surrounding environmental conditions, such as the temperature, the luminosity, the activity of potentially dangerous electric or gas devices, intruders alarm, etc. It can be specified as a tuple  $h = (ID_{home}, location, HomeAttr)$ . The set  $HomeAttr$

represents a set of attributes that adequately characterizes the user’s home. From now on, let us consider the set of  $H = \{h_1, h_2, \dots, h_n\}$ .

**Definition 3 (AAL Service)** - An AAL service can abstractly be defined as a tuple  $s = (IDservice, Description, SAttr)$ . The set  $SAttr$  represents a set of attributes, which adequately characterizes the AAL service. From now on, let us consider the set of  $S = \{s_1, s_2, \dots, s_n\}$ .

Recalling the recursive characterization of a service provider suggested in Fig. 11, a service provider  $sp_i$  may be a single entity or composed of other service providers. When it is composed of multiple entities, say  $sp_{i,1}, sp_{i,2} \dots$  and  $sp_{i,n}$ , their respective sets of services  $S_{i,1}, S_{i,2} \dots S_{i,n}$  can be combined in appropriate ways with an abstract composition operator, resulting in the composed services, as illustrated in Fig. 13. The set  $S_i$  is a subset of the all possible services composition  $2^{B_i}$ . Therefore,  $S_i$  holds the useful or profitable compositions only. The elements of  $S_i$  may in turn become basic services of other partnerships.

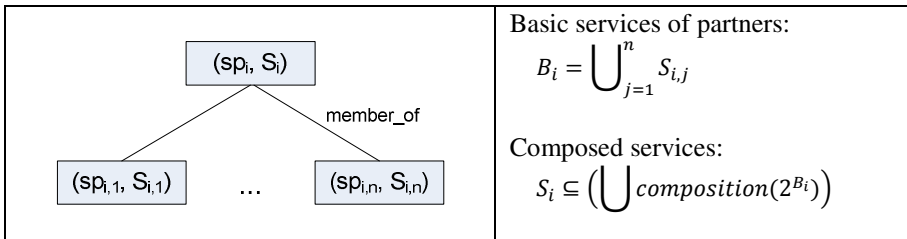


Fig. 13. A service provider as a composition of other service providers

**Definition 4 (AAL Service Provider)** - Entity that is able to provide care and assistance services to elders or persons who need assistance. They can be formal care providers, such as a hospital or a clinic; or informal care providers, such as churches or voluntary groups. One provider has a structure composed of one or several associated providers forming a partnership and can be recursively defined as a tuple  $pi = (\{p_{ij} \mid p_{ij} \neq p_i\}, S_i)$ , in which  $S_i \subseteq (\bigcup composition(2^{B_i}))$  and  $B_i = \bigcup_{j=1}^n S_{i,j}$ .

**Definition 5 (Subscription Contract)** - A subscription contract associates a user to a service provider by means of a number of services subscription. Each subscription can be specified as a tuple  $sc = (sp_i, u_j, S)$ . The set  $S = \{s_1, s_2, \dots, s_n\}$  represents the services subscribed by the user or delivered by the service provider.

Using these definitions, it is now possible to model mechanisms or rules to identify and select potential services that a user might subscribe.

**Definition 6 (Useful Services Selection)** - These are the services that a user or elder may need, according to its user attributes. The service selection can be formally obtained using the following query:

$$\forall_{s \in Services} \forall_{u \in Users} (UsefulService(s, u) \Leftarrow attributes(s, SA) \wedge attributes(u, SU) \wedge SA \cap SU \Leftarrow \emptyset).$$

In other words, the interception between service attributes and user attributes which are variables, results in a non-empty set.

In the last definition,  $SU$  and  $SA$  are free variables meaning they are not bound by universal or existential quantifiers. Using Definition 6, it is possible to select and propose services to a user, according to its characteristics or attributes, which can be a step before contract subscription. The mentioned definition can also be improved in order to consider the attributes of the user's home.

**Definition 7 (AAL Service Market Opportunity)** - Corresponds to an opportunity identified in the market, which may lead to the creation of a new tailored AAL service. It can be specified as a tuple  $MO = (ID, OPAttr)$ , in which  $OPAttr$  represents the set of attributes characterizing the opportunity. For now on, let us consider the set of market opportunities  $MO = \{mo_1, mo_2, \dots, mo_n\}$

**Definition 8 (AAL Service Provider Selection)** - Given a Market Opportunity  $mo$  (Definition 7), we can identify adequate partners through the following query:

$$\begin{aligned} \forall_{mo} \forall_{sp \in Providers} \forall_{criteria} (UsefulProvider(mo, sp) \\ \Leftarrow HasService(sp, s) \wedge attributes(s, SAttr) \\ \wedge attributes(mo, MOAttr) \wedge (SAttr \cap MOAttr) < \\ > \phi \wedge complies(sp, criteria)) \end{aligned}$$

The abstract operator ‘*complies*’ checks that candidate service providers do not clash with given business/strategic constraints.

**Definition 9 (AAL Partnership Formation)** - Given a market opportunity  $mo$ , either concrete or abstract, a corresponding partnership, or new service provider  $sp_i$  according to Definition 4, can be formed with the candidates service providers from the set  $\{sp_j \mid UsefulProvider(mo, sp_j)\}$ .

### 3.4 Data Models for the AAL Ecosystem

**The Global Infrastructure Model.** In general terms, the global infrastructure supports the interaction between the entities/nodes engaged in care provision, and the assisted people. It supports multi-node services, distributed processes, software services invocation and composition. The main functional blocks are: Global Infrastructure Management, Security Services, Software Services Composition, Safe Information Management Services at Global Level, Auditing Services, Safe Communication Service, and Auxiliary Services (including identification of critical issues, assessing performance, statistics, and reporting).

In our concrete specification, we assume that the global infrastructure takes care of a number of homes and users. Each home has got a number of ambient sensors



(e.g. temperature, blood pressure, etc.). Periodic observations are taken from these sensors, which might trigger important events. Some events may require an adequate response, like sending an SMS alarm to the user's relatives, or sending an emergency team to the user's home. The conceptual model presented in Fig. 14 incorporates these requirements.

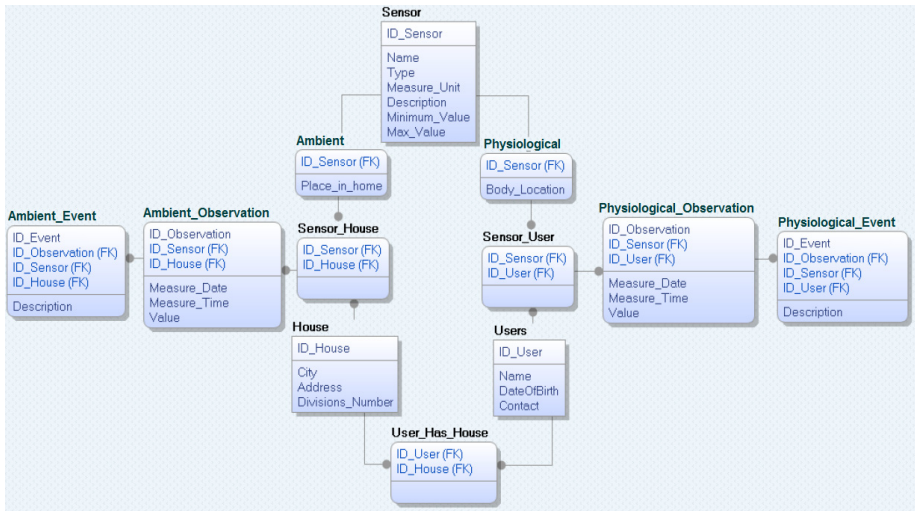


Fig. 14. Global infrastructure data model

**AAL Ecosystem Data Model.** In general terms, the top layer of the architecture – AAL Ecosystem – has the main purpose of providing, under a socio-technical perspective, organization and collaboration support for the AAL multi-stakeholders, organized as a collaborative community. Members of the AAL ecosystem include the AAL services/product providers, the end users, regulators, and other support entities such as governmental entities. This layer supports functionalities for establishing links between providers and users, business models, collaboration processes and governance policies enforcement. Main functional elements of this layer include: Ecosystem Management, Assets Management, Governance Policies Specification, Business Models Management, Providers User Linking Mechanisms, Consortia Formation Mechanisms, and Collaboration Support Mechanisms.

In our concrete specification, the ecosystem allows the creation of service providers, users, contract subscription, services and services composition, AAL events and billing. The conceptual model presented in Fig. 15 illustrates these aspects.

Whenever a user subscribes a service, a subscription contract between user and service provider is established. An AAL service can be a simple service or be recursively composed of other services, resulting in tailored packages. Such packages, for composed services, may be supplied by a single service provider. As established in

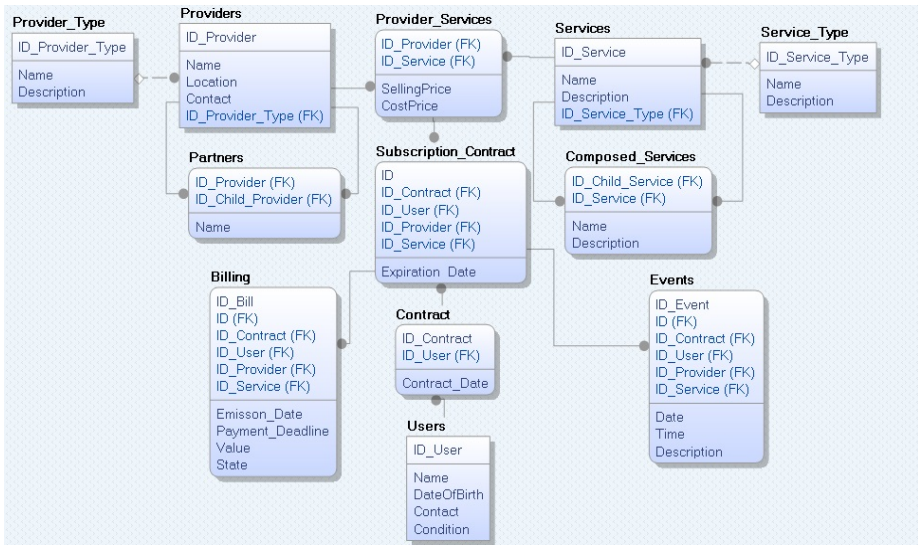


Fig. 15. Ecosystem management data model

previous section, for completing the package, the services of an additional partner may be necessary. In such case, the model allows the creation of providers that are in fact the combination of partners, resulting in partnerships. The Events entity encodes the main episodes related to these contracts, like the addition of more services to the contract. The Billing entity encodes the periodic payments resulting from the contract.

Having the specification established in previous section and the corresponding data models, we need to select an infrastructure for an adequate implementation, which is the subject of the next Section.

### 3.5 Technological Infrastructure

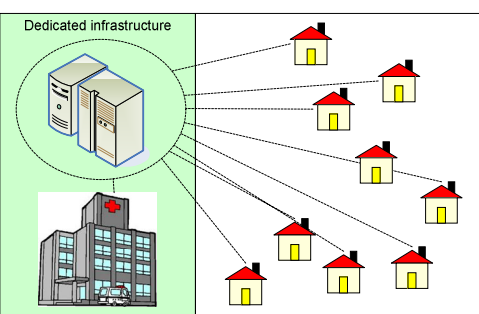
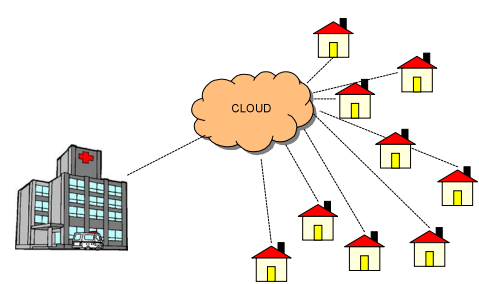
Ambient Assisted Living is a multifaceted area as it harnesses a diverse range of technologies from various domains. Requirements like remote supervision of elderly, information management and business processes, just to mention a few, are quite demanding in terms of ICT for the physical architecture of AAL. One of such necessities is the data centers which hold information of elders and monitoring data. The choice for adequate infrastructures is very important in terms of the necessary budget for launching an AAL business.

**The Cloud Computing Infrastructure.** Cloud computing environment is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. The cloud computing model provides essential characteristics, such as broad network access and resource pooling. It has got three service models, namely, Software as a Service, Platform as a Service and Infrastructure as a Service. It also provides sev-

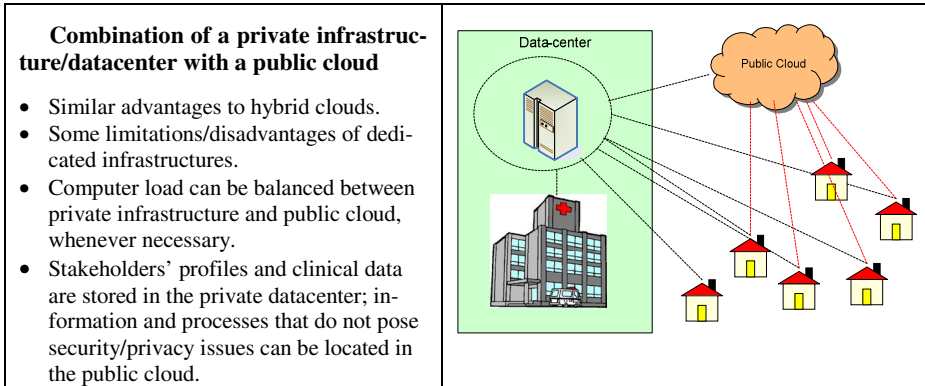
eral deployment models: Private Clouds, Community Clouds, Public Clouds, and Hybrid Clouds [32]. The typical features of Cloud Computing are, for instance, faster development/installation time, lower initial capital, and "pay-per-use".

Based on the possible deployment models, several strategies and infrastructures modalities for the development of the AAL4ALL ecosystem with Cloud Computing, together with their main features, are summarized in Table 5.

**Table 5.** Cloud computing modalities that are useful for the AAL ecosystem

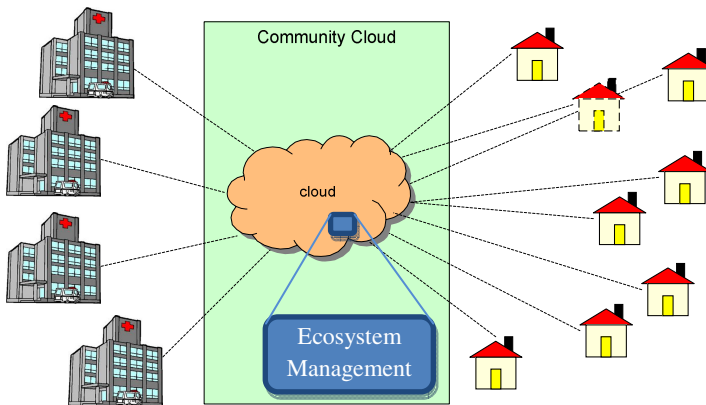
Modalities	Illustration
<p><b>Dedicated infrastructure/datacenter</b></p> <ul style="list-style-type: none"> <li>• High initial cost.</li> <li>• High development and launch time.</li> <li>• Complex installation and configuration.</li> <li>• The stakeholders are responsible for maintenance.</li> <li>• Some advantages are the complete ownership and control over the infrastructure, services and information.</li> </ul>	 <p>The diagram illustrates a dedicated infrastructure model. On the left, there is a server rack and a datacenter building. A dashed circle labeled 'Dedicated infrastructure' encloses the server rack. Lines connect the server rack to a group of ten houses, representing users, indicating that the infrastructure is exclusively for them.</p>
<p><b>Public cloud (provided by 3rd party supplier)</b></p> <ul style="list-style-type: none"> <li>• It is probably the fastest option to create and launch the ecosystem.</li> <li>• Many suppliers of Cloud Computing infrastructures are available.</li> <li>• Low initial cost. But if time is long, dedicated infrastructure may become cheaper.</li> <li>• Short Setup/installation time.</li> <li>• Cost is proportional ("Pay per use") to the number of users, utilization time, and used computational and storage resources, which might require some further analysis in face of the specific needs of AAL.</li> <li>• A major issue is that information of users (elders, customers and stakeholders), like the profiles and clinical information, is stored on third-party infrastructures, which raises security and privacy concerns.</li> <li>• An additional drawback is the lack of standards that allow portability. As such, once one provider is selected, solutions get too dependent on that provider.</li> </ul>	 <p>The diagram illustrates a public cloud model. On the left, there is a datacenter building. A central cloud icon labeled 'CLOUD' is connected to the datacenter building and to a group of ten houses, representing users. This indicates that the infrastructure is shared and provided by a third party.</p>

<p><b>Private cloud (owned by one stakeholder)</b></p> <ul style="list-style-type: none"> <li>• Has similar/same advantages of the public cloud.</li> <li>• Infrastructure is totally controlled by the stakeholder/AAL services provider.</li> <li>• Better in terms security and privacy, as management and access to information is performed by the stakeholder.</li> <li>• The stakeholder may start with a “small private cloud”, with a lower initial cost, and scale up the capacity if it becomes necessary afterwards.</li> <li>• This approach might suit a major services integrator.</li> </ul>	
<p><b>Community cloud</b></p> <ul style="list-style-type: none"> <li>• It has got similar advantages to private clouds.</li> <li>• Ecosystem acquires Cloud infrastructure to be shared by the stakeholders.</li> <li>• Ownership and control by the stakeholders of AAL4ALL ecosystem.</li> <li>• Acquisition, launch, and maintenance costs can be shared by the stakeholders which means lower costs and reduced business risks for each one.</li> <li>• Network installation and maintenance role can be assigned to a third-party provider, or rented/“outsourced” to a cloud-supplier, which already belongs to the ecosystem.</li> </ul>	
<p><b>Hybrid cloud</b></p> <ul style="list-style-type: none"> <li>• Similar characteristics of both public and private clouds.</li> <li>• Combination of public/community and private cloud infrastructures.</li> <li>• Computer load can be balanced between private and public clouds, whenever necessary.</li> <li>• Stakeholders’ profiles and clinical data are stored in the private cloud; information and processes that do not pose security/privacy issues can be located in the public cloud.</li> </ul>	



**The Ecosystem Web Portal.** From the available cloud-computing modalities, we developed the ecosystem portal in a public cloud (provided by third party supplier), as illustrated in Fig. 16. It was implemented using Microsoft Azure Cloud Computing [33].

A better choice for the Cloud Computing modality would be the community cloud described in Table 5. This modality allows benefiting from the cloud computing paradigm. But contrarily to a public cloud, it is owned by the stakeholders involved in AAL service provision. As a result, there would be fewer concerns in terms of the data and observations taken from the AAL users.



**Fig. 16.** Adopted cloud computing modality for the AAL ecosystem portal

The web portal allows the services providers to register and be part of the ecosystem. It allows partners to advertise their services to other partners and create partnerships whenever an opportunity arrives. Global infrastructures, which may be owned by a simple partner or by service provider's partnerships, are also implemented in the Cloud infrastructure. The main advantage is that infrastructures from providers can

scale as the number of users grow. In this way, each service provider can commit a lower initial budget and achieve reduced monthly cloud renting costs.

Fig. 17 illustrates the user interface regarding services subscription contracts established between service providers and users.

Fig. 17. Services subscription of a user

**The Local Infrastructure Nodes.** For the local nodes, representing the elders and their homes, an application that simulates a local infrastructure, which represent each user’s nodes a homes was designed. As illustrated in Fig. 18, several AAL services were subscribed, ranging from ambient to physiological monitoring. Each UI represents a user and its home. Whenever an event is triggered, in this simulation approach performed by a click in the UI control, the information is recorded in the corresponding table of the data model instantiated in the provider’s global infrastructure, illustrated at the right side of the mentioned figure.

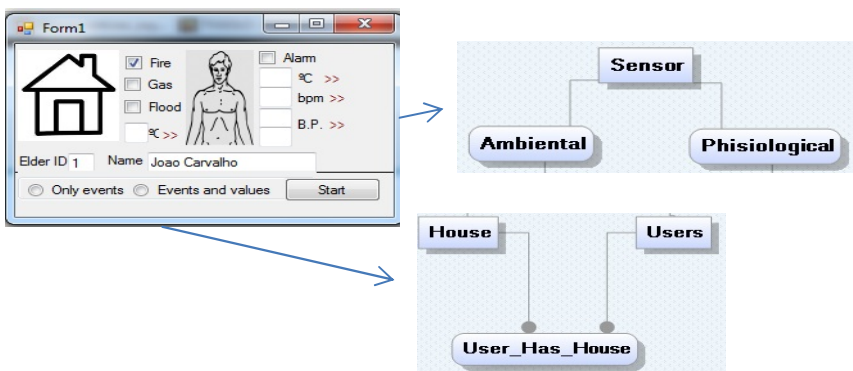


Fig. 18. Local infrastructure node (at the user’s home) and corresponding data model from global infrastructure

**The Global Infrastructure Nodes.** The global infrastructure node was also developed using a simulation approach, as illustrated in Fig. 14, which also illustrates the corresponding data model. Each observation taken from the local nodes is stored in the database running in the global node. The database complies with the models illustrated in Fig. 14 and Fig. 15. Such observations might trigger events that are relevant for the comfort and health of the users. These are the events shown in the user interface of the global node shown in Fig. 19.

The set of rules that identify events from observations are hard-encoded in C# inside the application. In this regard, we are planning the development of a rules based knowledge base, which provides more advanced events detection and corresponding reach at user’s homes. This is scheduled for future work.

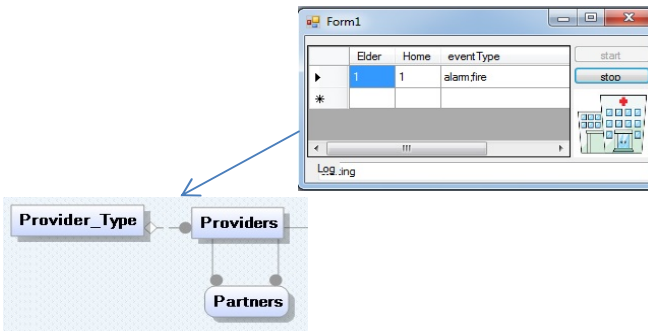


Fig. 19. Global infrastructure node and correspondent data model

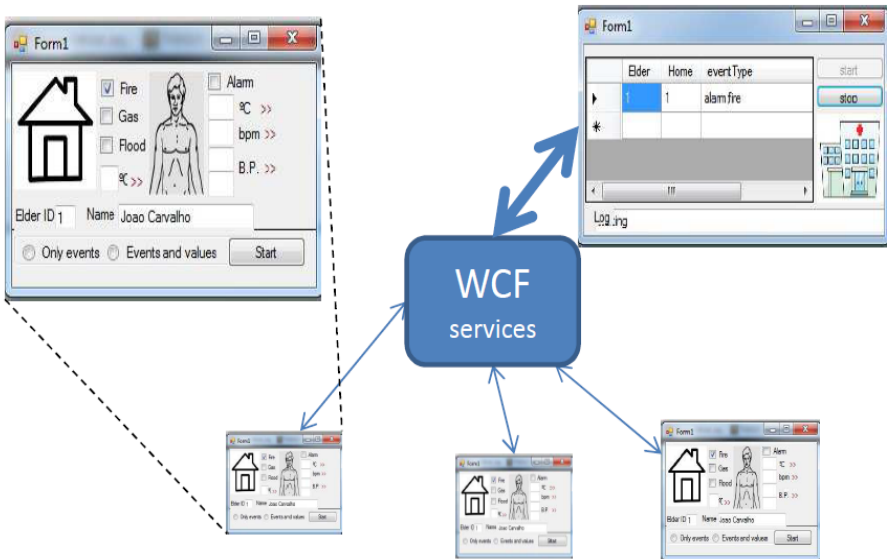


Fig. 20. Global and local infrastructures interacting through WCF

The interactions between the global and local nodes are illustrated in Fig. 20. These nodes exchange information regarding the mentioned events that are generated inside the local nodes, according to the subscribed AAL services. The mechanism for sending the events is based on REST services, which is supported by the Windows Communication Foundation (WCF) framework [34].

**The AAL System as a Whole.** The AAL Ecosystem platform was assembled together as a complete simulation system, in which information regarding AAL events flow from users and users' homes (local infrastructures) into global infrastructure nodes (Fig. 21). Member's management, services subscription, billing and other previously established requirements are fulfilled in the cloud portal, illustrated at the top of the mentioned figure. Partnership creation is also registered through the mentioned portal.

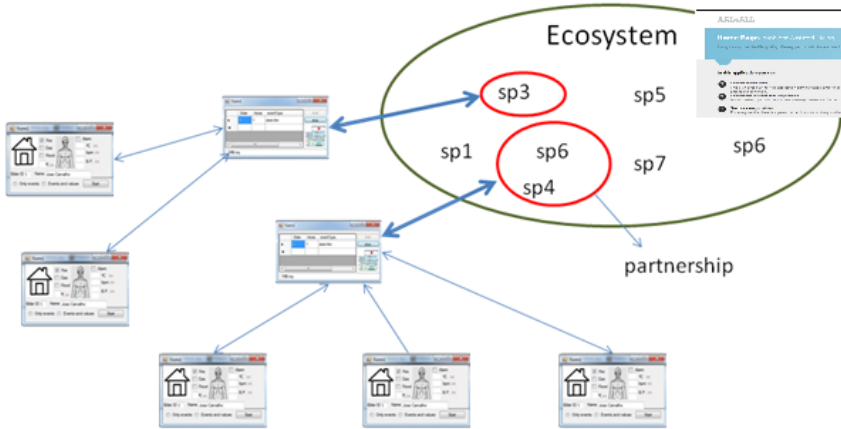


Fig. 21. The AAL ecosystem implemented as a simulation system

During the simulation of the entire system, we could verify and certify that implementation of the components follow the specifications established for the AAL ecosystem, and that it also fulfills the functional requirements that were established for each component of the ecosystem, namely, users, providers, services, and ecosystem management. Each time a new user is registered, the corresponding UI of the local node is created. A pre-specification number of AAL services is also launched.

While the complete system is operated, we can perceive its dynamics, in which each local node sends observations from the AAL services to the service providers' nodes. Information regarding these observations and events is stored in the corresponding tables of the ecosystem database.

**Analysis of the System.** Before ending the development description of the system, it is important to mention some remarks on the outcomes of the simulated ecosystem.

Although our approach for testing the ecosystem was based on simulation, the specification and data models were used as if the system was a real one. Simulation of



events flowed in real clock, in which we could simulate specific situations in the user's homes, which would allow service providers to trigger assistance services and home interventions. Furthermore, we are planning to integrate real nodes (e.g., real homes and users) in the simulation system, in the future work.

The strategy of simulating the AAL ecosystem can be seen as prior step to design a real ecosystem. That is because through the several performed simulations, we could find mistakes and improve specifications. Without simulation, we would have to fix the mistakes during system development or operation, which would cause the increase of the costs and potentially harmful effects on the users requiring remote monitoring and assistance services. In this regard, we take simulation as a system design paradigm [5].

## **4 Conclusions**

### **4.1 Synthesis of the Work**

During this research work, the implementation of an Ambient AAL Ecosystem was proposed, which uses technology as a way to improve the independence and wellbeing of aged people.

In order to define the best approach for such implementation, the first step was to review the state of the art, considering the technological aspects of an AAL structure, and its services and providers. With this literature review, we concluded that current approaches have been too techno-centric and realized that a collaboration-based approach would be more promising in terms of impacts in AAL area.

Our contribution to this effort was to specify and instantiate the ecosystem layer proposed by the AAL4ALL project. We started by identifying the functional requirements of the ecosystem management layer, followed by corresponding specifications and data models. These were formulated as canonical models, which allow the characterization of complex systems using simple, yet useful structures.

Afterwards, we developed the ecosystem management application in a Microsoft Azure Cloud Computing infrastructure. Local and Global nodes were implemented in C#. The interaction between these nodes was based on WCF.

The system was tested using a simulation approach, which allowed the understanding of the dynamics inside the ecosystem, certify the correctness of the specified models, fix both design and implementation mistakes, and perceive whether we could use our models for the development of real ecosystem.

### **4.2 Achieved Results**

During this work, we focused on the specification and implementation of concepts and structures for developing an AAL ecosystem. As mentioned before, we used canonical-based specifications to simulate the dynamics of the ecosystem life-cycle. With this purpose in mind, several aspects needed to be studied, in order to take the best possible decisions and to achieve satisfactory specifications and models. The results that were achieved include the following:

- Study and characterization of the AAL structure that was used in this project.
- Specification of functional requirements and specification of corresponding AAL architecture.
- Development of the ecosystem management system using a Cloud Computing framework.
- Development of local and global infrastructure nodes, which interact through WCF.
- Verification of ecosystem specification and its partial validation through simulation.

### 4.3 Future Work

Before starting to suggest future lines of action, it is worth to mention the context of this work. As mentioned before, the study and creation of an AAL ecosystem is a task which currently involves tens of researchers inside the AAL4ALL project. In this project, each one is working on concrete parts of the ecosystem development. This is important for our lines of future work, because it is recommended our future effort complements those at the project. As such, our strategy for future work is more focused on aspects, which will increase the functionality and quality of our AAL ecosystem, and at the same time, will profitably complement the tasks and results for the AAL4ALL project.

Therefore, our set of future work action comprises the integration of real nodes and users in the simulated ecosystem. The integration of these nodes in the simulated ecosystem, would allow the progressive transformation of our system from simulation to a real one. Additionally, this would then be installed in a number of homes, as a way to certify that the proposed ecosystem was specified in a way that allows further development towards real products, which may be “marketed”.

Other necessary and very useful component is the development of a knowledge-based system, which by making inferences with the observations taken from sensors, would trigger the corresponding events, from which service provider would provide corresponding assistance. For instance, this component would then select the adequate intervention regarding the event, sometimes sending an SMS to relatives, other times sending a rescuing team to the user’s home.

Other functionality for future work is the integration of business processes and service composition inside the simulations of the ecosystem. This would allow providing more complete AAL services, which would require several steps and several actors for their execution. Finally, we would like to incorporate in our system the capacity of modeling users with newly or emergent necessities. This would allow determining how to automatically formulate tailored packages of services and formation of corresponding partnerships.

**Acknowledgments.** This work was funded in part by the Project AAL4ALL (QREN 13852), co-financed by the European Community Fund through COMPETE - Programa Operacional Factores de Competitividade. The authors also thank the contributions from their partners in this project.

## Bibliography

1. Steg, H., Strese, H., Loroff, C., Hull, J., Schmidt, S.: Europe is facing a demographic challenge Ambient Assisted Living offers solutions. IST project report on ambient assisted living (2006)
2. Destatis. Older people in Germany and the EU. Wiesbaden: Federal Statistical Office of Germany (2011)
3. Camarinha-Matos, L.M., Afsarmanesh, H.: Collaborative Ecosystems in Ageing Support. In: Camarinha-Matos, L.M., Pereira-Klen, A., Afsarmanesh, H. (eds.) PRO-VE 2011. IFIP AICT, vol. 362, pp. 177–188. Springer, Boston (2011)
4. T113 - Arquitectura Tecnica AAL4ALL (2013), <http://www.aal4all.org> (seen on February 16, 2013)
5. Heilala, J., Vatanen, S., Tonteri, H., Montonen, J., Lind, S., Johansson, B., Stahre, S.: Simulation-based sustainable manufacturing system design. In: WSC 2008 -Winter Simulation Conference. IEEE (2008)
6. eNterface (2013), <http://eventos.fct.unl.pt/enterface13> (seen in September 5, 2013)
7. Eurostat, “Population statistics”, [http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/main\\_tables](http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/main_tables) (seen in December 12, 2013)
8. “Protégé Ontology Editor”, Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine (2013), <http://protege.stanford.edu> (seen in September 9, 2013)
9. Kutz, O., Mossakowski, T., Galinski, C., Lange, C.: Towards a standard for heterogeneous ontology integration and interoperability. In: International Conference on Terminology, Language and Content Resources (2011)
10. Becker, M.: Software architecture trends and promising technology for ambient assisted living systems. In: Proceedings of Assisted Living Systems—Models, Architectures and Engineering Approaches. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany (2008)
11. Camarinha-Matos, L.M., Rosas, J., Oliveira, A.I., Ferrada, F.: A Collaborative Services Ecosystem for Ambient Assisted Living. In: Camarinha-Matos, L.M., Xu, L., Afsarmanesh, H. (eds.) Collaborative Networks in the Internet of Services. IFIP AICT, vol. 380, pp. 117–127. Springer, Boston (2012)
12. Fernández-Llatas, C., Mocholi, J.B., Sanchez, C., Sala, P., Naranjo, J.C.: Process choreography for Interaction simulation in Ambient Assisted Living environments. In: XII Mediterranean Conference on Medical and Biological Engineering and Computing, pp. 757–760. Springer, Heidelberg (2010)
13. Menge, F.: Enterprise service bus. In: Free and Open Source Software Conference, vol. 2, pp. 1–6 (2007)
14. “T114 - Vigilância Tecnológica - Projectos Internacionais” (2012), <http://www.aal4all.org> (seen in September 4, 2013)
15. UbiSense (Ubiquitous Sensing and Behaviour Profiling), <http://www.ubicare.org/projects/subisense.shtml>, <http://www.doc.ic.ac.uk/vip/ubisense/> (seen in August 20, 2013)
16. ROSETTA, <http://www.aal-rosetta.eu/> (seen in October 10, 2010)
17. Dreaming (eDeRly-friEndly Alarm handling and MonitorING), <http://www.dreaming-project.org/> (seen in December 12, 2012)

18. ITALH (Information Technology for Assisted Living at Home), <http://www.eecs.berkeley.edu/~eklund/projects/ITALH/> (seen in April 5, 2013)
19. OASIS (Open architecture for Accessible Services Integration and Standardization), <http://www.oasis-project.eu/> (seen in July 1, 2013)
20. I2Home (Intuitive Interaction for Everyone with Home Appliances based on Industry Standards), <http://www.i2home.org> (seen in September 20, 2013)
21. AWARE (Ageing Workforce towards an Active Retirement), <http://aware.ibv.org/> (seen in August 30, 2013)
22. AALIANCE Project, <http://www.aaliance.eu/public/> (seen in April 12, 2013)
23. ePAL Project, <http://www.epal.eu.com/> (seen in June 5, 2013)
24. SENIOR Project, <http://www.seniorproject.eu/> (seen in June 10, 2013)
25. BRAID Project, <http://braidproject.org/> (seen in June 10, 2013)
26. Equivital, <http://www.equivital.co.uk> (seen in September 20, 2013)
27. Sensium, [http://www.toumaz.com/page.php?page=sensium\\_life\\_platform](http://www.toumaz.com/page.php?page=sensium_life_platform) (seen in October 20, 2013)
28. Hallo Monitoring, <http://www.halomonitoring.com/> (seen in December 12, 2013)
29. Grand Care, <http://www.tellaboomer.com/14.html> (seen in July 21, 2013)
30. HomMed, <http://www.hommed.com/> (seen in December 10, 2013)
31. Rosas, J., Camarinha-Matos, L.: A Collaboration Readiness Assessment Approach. *Innovation in Manufacturing Networks*, 77–86 (2008)
32. Mell, P., Grance, T.: The NIST definition of cloud computing (2011), <http://www.nist.gov/itl/cloud/> (seen in July 20, 2013)
33. Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolsvold, A., McKelvie, S., Xu, Y., Srivastav, S., Wu, J., Simitci, H.: Windows Azure Storage: A highly available cloud storage service with strong consistency. In: *Proceedings of the Twenty- Third ACM Symposium on Operating Systems Principles*. ACM (2011)
34. Mackey, A.: Windows Communication Foundation. *Introducing .NET 4.0*, pp. 159–173. Springer (2010)

# Author Index

- Ach, Laurent 50  
Al Moubayed, Samer 80  
Astrinaki, Maria 20
- Babacan, Onur 20  
Bantegnie, Emeline 50  
Barbulescu, Adela 20  
Baur, Tobias 50  
Ben Madhkour, Radhwan 179  
Berthouze, Nadia 50  
Beskow, Jonas 80  
Bollepalli, Bajibabu 80  
Bruijnes, Merijn 114
- Cakmak, Huseyin 20  
Camarinha-Matos, Luis M. 200  
Cardoso, Tiago 3, 141  
Carrasco, Gonalo 160  
Carvalho, Gonalo 200  
Coelho, Tiago 3
- d'Alessandro, Nicolas 20  
Dall, Rasmus 20  
Darriba Frederiks, Adu n 114  
Datta, Debajyoti 50  
de Oliveira, Rita 3  
Ding, Yu 50  
Dupont, St phane 50
- Ferrada, Filipa 200
- Gameiro, Jo o 141  
Griffin, Harry J. 50  
Grisard, Fabien 179
- Hu, Qiong 20  
Hueber, Thomas 20  
Huguenin, Victor 20  
Huisman, Gijs 114  
Hussen-Abdelaziz, Ahmed 80
- Johansson, Martin 80  
Jung, Merel 114
- Kalaycı, Emine S meyye 20  
Kliegr, Tomas 179  
Kolkmeier, Jan 114
- Koutsombogera, Maria 80  
Kuchar, Jaroslav 179
- Leroy, Julien 179  
Lingenfeller, Florian 50  
Lopes, Jos  David 80
- Mancas, Matei 179  
Mancini, Maurizio 50  
Moinet, Alexis 20  
Moreno, Alejandro 160
- Niewiadomski, Radoslaw 50  
Novikova, Jekaterina 80
- Oertel, Catharine 80  
Oliveira, Ana In s 200
- Parfait, Valentin 20  
Pelachaud, Catherine 50  
Pietquin, Olivier 50  
Piot, Bilal 50  
Pirner, Ivan 179  
Poppe, Ronald 160
- Ramos, Carlos 160  
Ravet, Thierry 20  
Reidsma, Dennis 160  
Rocca, Franois 179  
Rosas, Jo o 200  
Rybarczyk, Yves 3, 114, 141
- Skantze, Gabriel 80  
Stefanov, Kalin 80
- Tilmanne, Jo lle 20
- Urbain, J r me 50
- van Delden, Robby 160  
Varol, G l 80  
Vit, Jakub 179  
Volpe, Gualtiero 50
- Wagner, Johannes 50
- Zimmermann, Petr 179