Proceedings of eNTERFACE'16
The 12th Summer Workshop on Multimodal Interfaces

July 18th - August 12th, 2016, DesignLab, University of Twente
Enschede, The Netherlands

Organized by Human Media Interaction, University of Twente

Edited by
Khiet P. Truong & Dennis Reidsma

CTIT

CTIT Workshop Proceedings WP 17-02
ISSN 0929-0672

4TU.
HUMANS & TECHNOLOGY

UNIVERSITY OF TWENTE.

CTIT

Editors Khiet P. Truong & Dennis Reidsma
Enschede, the Netherlands (July, 2017)
CTIT Workshop Proceedings WP 17-02
ISSN 0929-0672

Workshop

Preface

The 12th Summer Workshop on Multimodal Interfaces eNTERFACE'16 was hosted by the Human Media Interaction group from the University of Twente (July 18th - August 12th, 2016). Four weeks long, students and researchers from all over the world came together in the DesignLab at the University of Twente to work on projects with topics evolving around the theme "multimodal interfaces". Universities and research institutes were invited to write project proposals that could be carried out during eNTERFACE'16. The following 9 projects were accepted and were present at the workshop (title of project and their principal investigators):

- A smell communication interface for affective systems (Emma Yann Zhang, Adrian Cheok)
- CARAMILLA: combining language learning and conversation in a relational agent (Emer Gilmartin, Benjamin R. Cowan, Nick Campbell, Ketong Su)
- Development of low-cost portable hand exoskeleton for assistive and rehabilitation purposes (Matteo Bianchi, Francesco Fanelli)
- Embodied conversational interfaces for the elderly user (Marieke Peeters, Mark Neerincx)
- Heterogeneous Multi-Modal Mixing for Fluent Multi-Party Human-Robot Interaction (Dennis Reidsma, Daniel Davison, Edwin Dertien)
- MOVACP: Monitoring computer Vision Applications in Cloud Platforms (Sidi Ahmed Mahmoudi, Fabian Lecron)
- SCE in HMI: Social Communicative Events in Human Machine Interactions (Hüseyin Çakmak, Kevin El Haddad)
- The Roberta IRONSIDE project: A dialog capable humanoid personal assistant in a wheelchair for dependent persons (Hugues Sansen, Maria Inés Torres, Kristiina Jokinen, Gérard Chollet, Dijana Petrovska-Delacretaz, Atta Badii, Stephan Schlögl, Nick Campbell)
- The Virtual Human Journalist (Michel Valstar)

In total, there were more than 65 participants coming from 16 different countries ranging from the USA to Malaysia and from the UK to Australia. We invited 3 keynote speakers to give a talk. Michel Valstar (University of Nottingham) presented about recent advances in computer vision (facial expression analysis) and machine learning. Kristiina Jokinen (University of Helsinki/University of Tartu) gave a talk about social engagement through eye-gaze in multimodal robot applications. And Anton Nijholt (University of Twente/Imagineering Institute) was invited to speak about smart technology and humor in playable cities. During the workshop, 3 courses were given by experts in the field. An introduction to deep learning for computer vision was taught by Gwenn Englebienne (University of Twente). Emer Gilmartin (Trinity College Dublin) led several informal academic English clinics. And a tutorial about the Social Signal Interpretation (SSI) framework was given by Johannes Wagner, Tobias Baur, and Dominik Schiller (University of Augsburg).

eNTERFACE'16 was made a successful event with the support of a whole team of people! We would like to thank the 4TU Centre on Humans & Technology and the CTIT for their financial support. The DesignLab, the Dream Team, Miriam and Erik: thank you for all your help and for staying open in the summer. Thank you to Jeroen, Randy, Daniel, Dennis, Dirk, Michiel, Charlotte, Alice, Wies, Lynn, and all who helped making this a great event! Thank you to the eNTERFACE steering committee for giving us the opportunity to host eNTERFACE'16.

Last but not least, we would like to thank all the participants in eNTERFACE'16. Thank you all for coming to Enschede and the DesignLab, it was great having you.

It was our pleasure hosting eNTERFACE'16 and we are looking forward to future eNTERFACE workshops.

Khiet Truong, Dennis Reidsma, Dirk Heylen
Human Media Interaction, University of Twente

July, 2017
Enschede, the Netherlands

List of participants

First name	Last name	Affiliation
Tuğçe	Akkoç	Bogazici University
Uğur	Ayvaz	Mugla Sitki Kocman University
Atta	Badii	University of Reading
Stephen	Barrass	University of Canberra
Tobias	Baur	University of Augsburg
Mohammed	Amine Belarbi	University of Mons
Matteo	Bianchi	University of Florence
Kevin	Bowden	University of California Santa Cruz
Tobias	Bützer	ETH Zurich
Sena	Büşra Yengeç	Turgut Ozal University
Angelo	Cafaro	Telecom ParisTech
Hüseyin	Çakmak	University of Mons
Oğuz	Çalık	Atılım University
Nick	Campbell	Trinity College Dublin
Kübra	Cengiz	Istanbul Technical University
Adrian	Cheok	Imagineering Institute
Gérard	Chollet	Intelligent Voice
Arianna	Cremoni	University of Florence
Daniel	Davison	University of Twente
Edwin	Dertien	University of Twente
Alpha Ousmane	Diallo	University of Trento
Mohammed	El Adoui	University of Mons
Kévin	El Haddad	University of Mons
Amr	El-Desoky Mousa	Technical University of Munich
Helena	Frijns	Leiden University
Alexandru	Ghitulescu	University of Nottingham
Emer	Gilmartin	Trinity College Dublin
Mahmut	Gökhan Turgut	Turgut Ozal University
Binnur	Görer	Bogazici University
Fasih	Haider	Trinity College Dublin
Mohammed	Hamoudi	University of Oran
Yuyun	Huang	Trinity College Dublin
Kristiina	Jokinen	University of Helsinki
Jaebok	Kim	University of Twente
Jan	Kolkmeier	University of Twente
Stefano	Laszlo Capitani	
Amine	Lazouni	University of Tlemcen
Minha	Lee	Eindhoven University of Technology
Jeroen	Linssen	University of Twente
Asier	López Zorrilla	University of the Basque Country
Sidi Ahmed	Mahmoudi	University of Mons
Siddharth	Mehrotra	International Institute of Information Technology
Seth	Montenegro	
Vivian	Motti	George Mason University
Mark	Neerincx	Delft University of Technology
Tommy	Nilsson	University of Nottingham

Javier Mikel	Olaso	University of the Basque Country
Marieke	Peeters	Delft University of Technology
Gueorgui	Pironkov	University of Mons
Blaise	Potard	Cereproc
Ahmed	Ratni	
Dennis	Reidsma	University of Twente
Hugues	Sansen	SHANKAA
Emre	Saraçoğlu	Turgut Ozal University
Bob	Schadenberg	University of Twente
Dominik	Schiller	University of Augsburg
Stephan	Schlögl	MCI Management Center
Nicola	Secciani	University of Florence
Omar	Seddati	University of Mons
Christine	Spencer	Queen's University Belfast
Ketong	Su	Trinity College Dublin
Mariët	Theune	University of Twente
María Inés	Torres	Universidad del Pais Vasco
Trung Ngo	Trong	University of Eastern Finland
Matteo	Venturi	University of Florence
Bob	Vijver, van de	University of Twente
Johannes	Wagner	University of Augsburg
Jelte	Waterschoot, van	University of Twente
Emma	Zhang	City University London
Yong	Zhao	Vrije Universiteit Brussel

,

Things that Make Robots Go HMMM: Heterogeneous Multilevel Multimodal Mixing to Realise Fluent, Multiparty, Human-Robot Interaction

Daniel Davison[1] Binnur Görer[2] Jan Kolkmeier[1] Jeroen Linssen[1] Bob Schadenberg[1] Bob van de Vijver[1,4]
Nick Campbell[3] Edwin Dertien[1] Dennis Reidsma[1]

***Abstract*—Fluent, multi-party, human-robot interaction calls for the mixing of deliberate conversational behaviour and reactive, semi-autonomous behaviour. In this project, we worked on a novel, state-of-the-art setup for realising such interactions. We approach this challenge from two sides. On the one hand, a dialogue manager requests deliberative behaviour and setting parameters on ongoing (semi)autonomous behaviour. On the other hand, robot control software needs to translate and mix these deliberative and bottom-up behaviours into consistent and coherent motion. The two need to collaborate to create behaviour that is fluent, naturally varied, and well-integrated. The resulting challenge is that, at the same time, this behaviour needs to conform to both high level requirements and to content and timing that are set by the dialogue manager. We tackled this challenge by designing a framework which can mix these two types of behaviour, using *AsapRealizer*, a Behaviour Markup Language realiser. We call this Heterogeneous Multilevel Multimodal Mixing (HMMM). Our framework is showcased in a scenario which revolves around a robot receptionist which is able to interact with multiple users.**

***Index Terms*—Social robotics, human-robot interaction, multi-party interaction, multi-modal interaction, Behaviour Markup Language.**

I. Introduction

THE main objective of this project is to bring forward the state of the art in fluent human-robot dialogue by improving the integration between deliberative and (semi)autonomous behaviour control. The interaction setting in which this has been done is one of multi-party interaction between one robot and several humans. The project builds upon interaction scenarios with collaborative educational tasks, as used in the context of the EU EASEL project [2], and uses and extends the state-of-the-art BML realiser *AsapRealizer* [3]. Fluent interaction plays an important role in effective human-robot teamwork [4], [5]. A robot should be able to react to a human's current actions, to anticipate the user's next action and pro-actively adjust its behaviour accordingly. Factors such as interpredictability and common ground are required for establishing such an alignment [6], [7]. Regulation of (shared) attention, which to a large extent builds upon using the right gaze and head behaviours [8], plays an important role in maintaining the common ground. In a multi-party setting, the matter becomes more complex. A mixture of conversational behaviours directed at the main interaction partner, behaviours directed at other people nearby to keep them included in the conversation, and behaviours that show general awareness of the surrounding people and environment need to be seamlessly mixed and fluently coordinated to each other and to actions and utterances of others.

For a robot that is designed to be used in such a social conversational context, the exact control of its motion capabilities is determined on multiple levels. The autonomous level controls behaviours such as idle motions and breathing. Secondly, the semi-autonomous level governs behaviours such as the motions required to keep the gaze focused on a certain target. Thirdly, there is a level for reactive behaviours such as reflex responses to visual input. Finally, the top level consists of deliberative behaviours such as speech or head gestures that make up the utterances of the conversation. Part of the expressions, especially the deliberative ones, are triggered by requests from a dialogue manager. Other parts may be more effectively carried out by modules running in the robot hardware itself. This is especially true for modules that require high frequency feedback loops such as tracking objects with gaze or making a gesture towards a moving object.

A dialogue manager for social dialogue orchestrates the progress of the social conversation between human and robot. Based on this progress, the manager requests certain deliberative behaviours to be executed and certain changes to be made to parameters of the autonomous behaviour of the robot. Such requests are typically specified using a high level behaviour script language such as the Behaviour Markup Language (BML), which is agnostic of the details of the robot platform and its controls and capabilities for autonomous behaviours [9]. The BML scripts are then communicated to the robot platform by a Behaviour Realiser (in this project: *AsapRealizer* [10]),which interprets the BML in terms of the available controls of the robotic embodiment. Behaviours, both autonomous and semi-autonomous, may then be mixed into the deliberative behaviours, either by AsapRealizer or by the robot platform itself. Since the behaviour should respond fluently to changes in the environment, the dialogue models as well as the robot control mechanisms must be able to adapt on-the-fly, always being ready to change on a moment's notice. Any running behaviour could be altered, interrupted or cancelled by any of the control mechanisms to ensure the responsive nature of the interaction. This multi-level control can include social

[1] University of Twente, Enschede, The Netherlands
[2] Boğaziçi University, Istanbul, Turkey
[3] Trinity College Dublin, Dublin, Ireland
[4] Part of the work on behaviour mixing was previously reported in [1].

commands like maintaining eye contact during conversations, as well as reactive commands like looking at sudden visually salient movements.

In this project we worked on such seamless integration of deliberative, (semi)autonomous behaviours for a social robot. This introduced a challenge for an architecture for human robot interaction. On the one hand, the robot embodiment continuously carries out its autonomous and reactive behaviour patterns. The parameters of these may be modified on the fly based on requests by the dialogue manager. On the other hand, the dialogue manager may request deliberative behaviours that actually conflict with these autonomous behaviours, since the dialogue manager does not know the exact current state of the autonomous behaviours. The control architecture therefore contains intelligence to prioritise, balance and mix these multilevel requests before translating them to direct robot controls. We call this Heterogenous Multilevel Multimodal Mixing (HMMM). In addition, the robotic embodiment sends updates and predictions about the (expected) timing with which behaviour requests from the dialog manager will be carried out, so the dialog manager can manage adaptive dialogue [11]. We intended to create a proof-of-concept system, showing how the required modules support multi-party interactions with behaviour-mixing. The resulting system has been showcased in a context in which fluent and responsive behaviour are shown off to good advantage. To this end we have set up a robot receptionist scenario centred around multi-party interaction with dynamic and responsive gaze behaviour.

The remainder of this paper is structured as follows. In Section II, we address work related to HMMM. We outline the scenario we chose to showcase our approach in Section III. Section IV describes the architecture of our system. This is followed up in Section V with the requirements of our approach. In Section VI, we describe the implementation of HMMM. We present our conclusions in Section VII.

II. Related Work

Many approaches to designing, implementing and evaluating social robots exist, see [12]–[14]. As explained in the introduction, for HMMM, we specifically looked at how to realise fluent, multi-party, human-robot interaction. In this section, we provide a high-level overview of existing work related to the different facets of our work.

According to Bohus and Horitz, the challenges for open-world dialogues with both robots and virtual agents originate in their dynamic, multi-party nature, and their situatedness in the physical world [15]. Bohus and Horitz address these challenges by developing a system with four core competencies: situational awareness through computer vision; estimation of engagement of users; multi-party turn-taking; and determination of users' intentions. In later work, Bohus and Horitz also address the difficulties of multi-party interaction, emphasising the importance of fluent turn-taking by stating that failure to do so leads to a system shifting 'from a collaborating *participant* into a distant and uncoordinated *appliance*' [16]. In his review of verbal and non-verbal human-robot communication, Mavridis proposes a list of desiderata for this field's state-of-the-art [13]. This list supports the importance of the challenges addressed by Bohus and Horitz, but also emphasises the necessity of affective interactions, synchronicity of verbal and non-verbal behaviour, and mixed-initiative dialogue. Whereas Mavridis focuses on requirements for interpersonal behaviour, functional open-world dialogues also require correct *intrapersonal* behaviour. Similar to work in the field of semantics on interaction cycles, such as that of [17] and implemented in a human-robot interaction system by [18], we model our interactions using a *sense-think-act* cycle. We enable our robots to gather data using different sensors, process and interpret this information, and finally carry out appropriate actions. In this paper, we address this necessity of the mixing of behaviour that is generated top-down and bottom-up, as argued in the introduction. Our approach builds on the challenges Bohusand Horitz erected as pillars of the field of human-robot dialogues in the wild.

Gaze behaviours can be utilised by a robot to shape engagement and facilitate multi-party turn taking [19]. In a conversation, gaze behaviours serve various important functions, such as enabling speakers to signal conversational roles, such as speaker, addressee and side participant [20], facilitating turn-taking, and providing information on the structure of the speaker's discourse [21]. Endowing robots with the capacity to direct their gaze at the appropriate interlocutor combined with the capability of doing this with the correct timing leads to more fluent conversations [22] and improves the interlocutors' evaluation of the robot [23].

In conversations with multiple interlocutors, it is important that the robot can accommodate to the various conversational roles, and the shifting of these roles over the course of the conversation. It should be clear to the interlocutors who the robot is addressing. In multi-party interaction between humans, a speaker's gaze behaviour can signal whom the speaker is addressing and whom is considered a side participant of the conversation [24]. The shifting of roles during conversation is accomplished through turn-taking mechanisms. For example, the addressee whom the speaker looks at the end of a remark is more likely to take up the role of speaker afterwards. In turn, by looking at the speaker at the end of the speaker's turn, an addressee can signal that he or she can take over the turn. For example, Mutlu et al. [19] found that a robot can also utilise gaze behaviours to successfully cue conversational roles in human participants. This can also be used to create equal engagement of multiple partners in a conversation [25].

Gaze behaviours are partly (semi)autonomous, but can also be used for deliberate and reactive behaviour. For instance, a deliberate use of gaze is when you direct your gaze to a cookie and stare intently at it, to communicate that you desire the cookie, whereas directing your gaze in reaction to a salient event that happened in your vicinity, is an example of a reactive use of gaze. For this project we therefore chose to focus on gaze behaviour as one of the modalities for exploring heterogeneous multi-modal mixing. As explained in the introduction, *AsapRealizer* already provides the necessary functionality to incorporate gaze behaviour based on deliberate and (semi)autonomous behaviour [3].

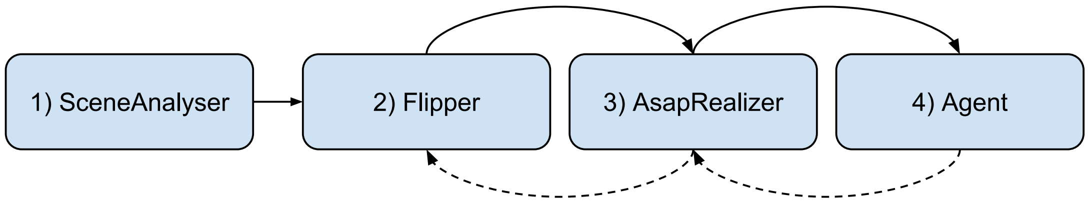


Fig. 1: An overview of the eNTERFACE system architecture, highlighting four distinct components: (1) the signal acquisition module SceneAnalyzer; (2) the dialogue manager Flipper; (3) the behaviour realiser AsapRealizer; (4) the agent (for example, the Zeno or EyePi robots, or a virtual agent created in Unity).

III. Scenario

We chose to let our human-robot interactions take place in a real-life context with the robot being a receptionist for a doctor's appointment. Users are given the goal of visiting one of two available doctors. The robot should be able to draw users' attention, welcome them, instruct them on which way to go to visit their doctor, and bid them farewell. When another user enters the detection range of the robot during its interaction with the first user, it should be able to recognise and acknowledge the second, possibly shifting its attention to that person. This conversational setting satisfies the prerequisites for multi-party capabilities, fluent behaviour generation, and mixing of deliberate and autonomous behaviour.

Our first working prototype incorporated a scenario for the receptionist robot interacting with a single user. The dialogue with the robot revolves around users having a goal of visiting one of two doctors. The robot assists users in finding their way to their appointments. Below, we discuss the conversational phases of the dialogue users can have with the robot. Appendix A discusses our work on the dialogue part of this project in more detail and shows the setup of the interaction (see Fig. A.1).

We chose to let the robot take initiative during the largest part of the interaction, letting it guide users through the dialogue in order to limit their agency and keep the interaction straight and simple. Building on our ideas of a suitable scenario for HMMM, our scenario consists of several *conversational phases*: *Initialise*, *Welcome*, *Instruct*, *Direct*, *Farewell*. These phases follow each other sequentially. In the first phase, the system is initialised and parameters are set. When a user enters the interaction range of the robot, the *Welcome* phase is started: the robot acknowledges the user with a short gaze and, if the user approaches even closer, the robot will say 'Hi!' to welcome her. During the *Instruct* phase, the robot will instruct the user to point at one of two nameplates showing the doctor she wants to visit. It does so by uttering the sentences 'Please point at the sign with your doctor's name on it. Either this sign on your left, or this sign on your right.' and by gazing and pointing at both nameplates in sync with this verbal utterance. If the user does not seem to comply with these instructions, the robot will try to instruct her again. If this fails again, the robot directs her to a nearby human for further assistance. When the user has pointed at the nameplate of a doctor, the dialogue enters the *Direct* phase. In this phase, the robot directs the user in the correct direction for her doctor, again talking, gazing and pointing. Similar to the previous phase, if the user does not seem to understand these directions, the robot will direct her again before finally directing her to a nearby human if she fails to respond for a third time. If it turns out that the user walks off in the wrong direction after the robot's directions, it will call her back, directing her once more in the correct direction. Finally, when the user walks off in the correct direction, the robot offers her a friendly smile and waves at her, saying 'Goodbye!' in the *Farewell* phase. Thereafter, the system returns to the *Initialise* phase, ready for a new user.

IV. Global Architecture

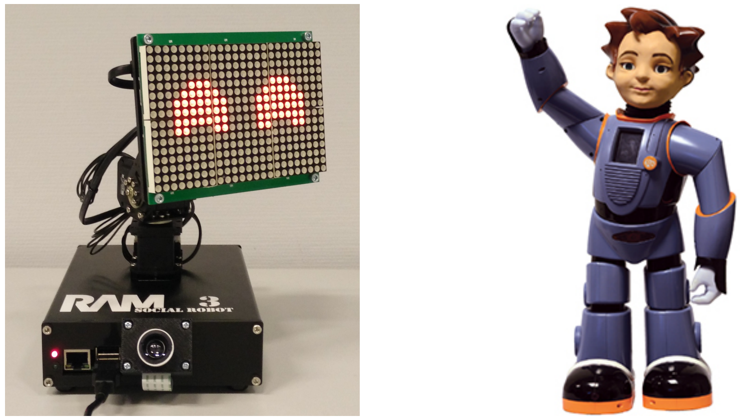
The fluent behaviour generation system described in this report uses a layered modular architecture. This architecture is designed to separate the various processes required for generating appropriate behaviour into standalone components, which communicate through a common middleware. A more comprehensive version of such an architecture is described by Vouloutsi et al. in the EASEL project [26]. In this section we present a streamlined version of the architecture developed for eNTERFACE '16, focusing primarily on the components involved in generating fluent dialogues and behaviour. The global architecture consists of four components, see Fig. 1.

A. Perception

The perception module provides information about the state of the world and the actions of an interlocutor. Such information is crucial for making informed decisions about which appropriate behaviours to execute in the current state of the dialogue. The *SceneAnalyzer* [27] application uses a *Kinect* sensor to detect persons in interaction distance. Amongst other data, it estimates the probability that a person is speaking, and it extracts the location of the person's head, spine and hands. This data is further processed to extract features such as proxemics and (hand) gestures.

B. Dialogue Manager

The role of the *Flipper* [28] dialogue manager component is to specify, monitor, and manage the flow of a dialogue. By interpreting actions of the user, and taking into account the current context of the interaction, the dialogue manager selects an appropriate *behavioural intent* to convey to the user. This behavioural intent is then translated to BML behaviour



(a) The EyePi robot. (b) The Zeno R25 robot.

Fig. 2: The robotic platforms used during the project.

commands suitable for an agent embodiment. This two-step abstraction of *interaction context* to *behavioural intent* to BML allows us to define a high-level flow of a dialogue that is independent from the low-level agent controls. More details of the high-level dialogue implementation are given in Section III and Appendix A. Using this method, a high-level dialogue will be able to generate behaviour for any agent platform, as long as the agent is able to express the behavioural intents using its platform-specific modalities.

C. Behaviour Realiser

The *AsapRealizer* is a BML behaviour realiser engine, that takes behaviour specifications and translates these to agent-specific control primitives [10], [29]. The realiser is capable of resolving inter-behaviour synchronisation, resulting in a detailed schedule of planned behaviour fragments, such as speech, gaze and animations [11]. These behaviour fragments are mapped to a agent-specific control primitives, each of which might have different timing constraints. Such control primitives can include joint- or motor-rotations, text-to-speech requests, or animation sequences. To determine the exact timings for a specific agent, AsapRealizer relies on a negotiation process with the agent embodiment. During execution of the behaviours, AsapRealizer receives feedback from the agent embodiment about the progress of execution, which is necessary for planning on-the-fly interruptions and adaptations of the planned behaviour [11]. This process is described in more detail in Section VI-A.

D. Agent Control

We focused on controlling two specific robot agents: *EyePi* (Fig. 2a) and *Zeno R25* (Fig. 2b). The EyePi is a minimalistic representation of a robotic head and eyes, offering fluent control over gaze direction, emotional expressions, and a collection of animation sequences. It has three autonomous behaviours: breathing, done by rhythmically moving up and down its head; blinking its eyes; and gazing at salient points in its field of view (see Section VI-C). The Zeno R25 is a small humanoid robot, offering control over gaze direction, facial expressions, animations and speech. Whereas the EyePi has very responsive and fluent control, the Zeno R25 offers more channels to express visual modality, such as hands and a fully expressive face. However, it only has one autonomous behaviour, namely blinking.

We used these two embodiments to showcase how they can be controlled from one general architecture and how to enable multi-party interaction with them. The process of extending AsapRealizer with new embodiments based on these control primitives is described in more detail by Reidsma et al. [10].

V. Requirements

Within the context of the 'pillars' of Bohus & Horitz [15] as discussed in Section II, we constructed a demonstration of a fluent multi-party interaction. In this section we describe the specific, additional requirements for achieving our global aim. These revolve around three main themes: fluent behaviour generation (Section V-A), multi-party capabilities (Section V-B), and behaviour mixing (Section V-C).

A. Fluent Behaviour Generation

Generating behaviour that can adapt fluently to external influences introduces several requirements for our system architecture. Figure 3 shows an abstract overview of the behaviour generation pipeline, consisting of a dialogue manager, a behaviour realiser and one or more agent control engines. Generally, the dialogue manager runs a high-level dialogue model, that specifies how an agent should respond when interacting with a user. The dialogue model sends BML behaviours to a behaviour realiser (1), which then translates these to agent-specific commands. These low-level commands are sent to an agent control engine (2), which executes the actual behaviour on an agent embodiment (for instance, a virtual human or a robot). Welbergen et al. give a detailed explanation of the processes required for performing behaviour realisation on an agent [29]. In Section IV, we give an architectural overview of our dialogue manager, behaviour realiser and agent control engines.

Not all agents are identical in the way they handle behaviour requests. Typically, a virtual character offers very predictable

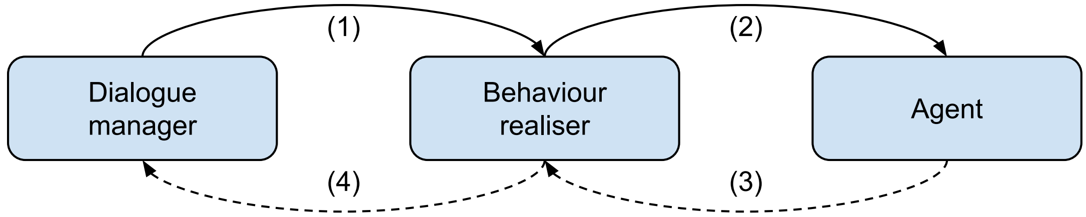


Fig. 3: Abstract overview of the fluent behaviour generation pipeline: (1) the dialogue manager generates BML behaviour; (2) the behaviour realiser generates agent-specific commands; (3) the agent delivers feedback about planning and execution of these commands; (4) the behaviour realiser provides feedback about the behaviour progress.

controls in terms of motion and timing. For example, gazing at a 'red ball' object in 0.2 seconds will be executed without much problem by the 3D rendering engine. However, a physically embodied agent, such as a robot, might have physical limitations to its movements which makes it more difficult to accurately predict its movements and timings. For example, gazing at the 'red ball' in 0.2 seconds might be physically impossible due to limitations in the actuators. Depending on what the current gaze direction is, it could instead take 0.5 seconds. This delay needs to be communicated back to the behaviour realiser to ensure correct synchronisation with other planned behaviours. Additionally, dynamic environmental factors such as temperature or battery level might play a role in predicting and executing physical behaviours.

Concretely, this means that the behaviour realiser not only needs to negotiate in advance with the agent control engine about expected timing of certain gestures and actions, but that it also needs to be kept up to date about actual execution progress. This way, it can adapt the timing of other, related, behaviours such as speech. Specifically, feedback from the agent control engine about command planning and execution (3) is required to perform inter-behaviour synchronisation. Feedback from the behaviour realiser about BML behaviour progress (4) is used to perform dialogue synchronisation and validation. This is discussed in more detail in [29].

Specifying and implementing adequate feedback mechanisms are important requirements for fluent behaviour generation and adaptation, on both the dialogue level and the behaviour realisation level. In Section VI we discuss our approach and give several examples where this is used to generate more fluent behaviour patterns.

B. Multiparty Capabilities

An interaction with a user often does not take place in an isolated, controlled environment. There is always a possibility for distractions or interruptions, which might require an agent to adapt its running or scheduled behaviour. Resynchronising, rescheduling and interrupting individual behaviours is typically handled by the behaviour realiser. However the *decision* to perform these behaviour modification actions is driven by the agent's dialogue model, based on an interpretation of the environment and the current interaction: 'Is there an interruption that is relevant? What am I doing at the moment? Does it make sense to stop what I am doing and do something else instead?'

Assuming that we have an agent control architecture that can perform fluent behaviour generation, as described in the previous section, we can use the feedback about behaviour progress to plan and execute behaviour interrupts and reschedule future behaviours on a dialogue level. We use this functionality to incorporate multiparty capabilities in a dialogue. For a fluent integration of other interlocutors in an interaction, the multiparty capabilities should include: (1) tracking of multiple interlocutors; (2) acknowledgement of each (new) interlocutor, well-coordinated with the ongoing interaction with the main interlocutor; (3) assessment of each interlocutor's priority for gaining the focus of attention; (4) dialogue mechanisms for interrupting and switching between interlocutors.

C. Behaviour Mixing

The final main requirement for our system concerns the Heterogeneous Multilevel Multimodal Mixing, the necessity of which we argued in the introduction. Autonomous behaviour, such as breathing motions, eye blinking and temporarily gazing at interesting objects, must be combined with deliberate behaviours in a seamless way. We focus on head behaviours as a use case for different types of behaviour mixing. More specifically, we look at three types of head behaviour: gaze direction, based on a combination of visual saliency maps; emotional expressions, based on valence/arousal space; and head gestures such as nodding, shaking, or deictive gaze (pointing at an object using the head). Any robot platform that implements these high level behaviours can be controlled in a transparent manner by AsapRealizer. We focus on the EyePi as a platform that additionally can mix conflicting or complementing requests before actually executing them. Section VI describes how we implemented these capabilities.

VI. Implementation

In order to achieve fluent, multi-party human-robot interaction, we extended *AsapRealizer*, and implemented a design pattern in *Flipper*. In this section we first describe how we implemented fluent robot control, followed by the design pattern through which we achieved multi-party interaction. The remaining subsections describe how we mix various behavioural modalities.

A. Fluent Robot Control

To realise fluent behaviour for our robots, we implemented feedback mechanisms between them and *AsapRealizer*. This involved aligning the control primitives of both robot platforms with *AsapRealizer*'s BML commands and Flipper's intents as incorporated in the dialogue templates. Feedback is provided on several levels: (1) feedback on whether the behaviour has been performed or an estimation of its duration; (2) an estimation of its duration before execution, with real-time updates when running; (3) a combination of the former two, including real-time adjustment of running synchronisation points. For further detail on these levels of feedback, we refer to [29]. We implemented these feedback mechanisms in the EyePi and Zeno platforms.

Specifically, for the EyePi platform, these are the following:

1) *It is impossible to plan this request* **(nack)**
   It is not possible to execute the requested sequence. This might be the case if the the requested time is to soon or has already past, or the actuators are not available.
2) *Exact negotiation* **(–)**
   This feedback type will be used when the requester wants to know when a specific sequence can be planned such that it will be executed. The requester will need to send a new request with the required timing based on the negotiation result.
3) *Negotiation* **(ack)**
   This feedback will be used if the requester specified that the sequence should have a start on or after the requested time. This is a weak request, on which the feedback will contain the computed planning.
4) *Try to execute, but motion parameters are updated* **(ack)**
   If it is possible to achieve the timing by updating the motion parameters (within configured bounds), the parameters will be updated and will also be send as feedback.
5) *Will execute, but it will be late* **(ack)**
   If the requested timing can not be met, but it can be met if it is within the configured flexibility limit (for example, 50-100 ms), the sequence will be executed.
6) *Will execute on time* **(ack)**
   If the requested timing can be met without problems.

The timing requests can be made on the start, stroke and end synchronisation points and that all feedback holds for every possibility. To be able to handle the stroke and end timing requests, the sequence mixer calculates the expected duration of every sequence request from arrival as they are based on the currently active motion parameters.

B. Multiparty Interaction

In order to accommodate interrupts from a bystander, we developed general patterns for the dialogue management scripts that are independent of the actual contents of the ongoing dialogue. We implemented a priority system that signals the importance of the current discourse, and any event that may occur during a conversation. A priority, ranging from 1 (low) to 3 (high), is assigned to each dialogue template in Flipper (see Fig. A.2 and Appendix A), which defines the importance of the continuity of the behaviour that is linked to the template. For example, when the robot is giving directions to the addressee, an interruption would severely disrupt the interaction. Therefore, the dialogues in which the robot gives directions are given a high priority. Behaviours generated as part of this should not be interrupted for the sake of relatively unimportant additional events. When the robot has completed an action, the priority threshold is lowered again.

Next to the dialogues, each bystander which is recognised in the scene also receives a priority. When a bystander is recognised for the first time, a low priority is assigned to the bystander. The priority is increased when the bystander actively tries to get the attention of the robot by either talking or waving with the arms. When either is recognised, the bystander's priority will increase to a medium priority. When the bystander is both talking and waving, he or she is given a high priority.

Whether or not the robot responds to the bystander depends on matching the detected priority of the bystander with the dialogue's priority, as defined in the Flipper templates. When the bystander's detected priority is smaller than the priority of the dialogue, the bystander will be ignored. In such cases, the agent will continue its current discourse with the main interlocutor. If the bystander's priority becomes equal or larger than the dialogue's priority, the agent's attention will shift towards the bystander.

When the robot responds to a bystander, the actual form of this response is determined by the priority of the bystander. When the bystander has a low priority, the robot switches its gaze to the bystander to acknowledge their presence, and then returns its gaze to the interlocutor. In case the bystander has a medium priority, the robot will address the bystander by gazing at the bystander and telling him or her to wait. Alternatively, when the bystander has a high priority, the robot will tell the main interlocutor to wait, and start a conversation with the bystander; the main interlocutor and bystander switch roles. After finishing the conversation with the new interlocutor, the robot will continue with the conversation that was put on hold.

C. Behaviour Mixing

In our system, the behaviour mixing is divided into three parts: emotion, gaze and sequence. All generate an output, which is being handled by the robot animator that converts those command directly into movement. While the animator is robot specific, the mixing can be reused for every robot that wants to support emotion, gaze and sequence commands.

1) Emotion Mixing: The emotion mixing part of HMMM (Fig. B.1) can be considered the simplest mixing part. It processes input from both external requests and requests from the gaze component. Emotion requests are composed of a valence and arousal parameter, as defined by Russell's circumplex model of affect [30]. The requests are directly mixed and new output values are calculated based on the current state and the requested values.

In the current implementation, external emotion requests are composed of deliberate emotions that accompany certain behaviour or speech, and have been pre-crafted by the dialogue designer. Other emotion requests could originate from sources such as: automatic responses based on emotions detected from the user; or personality and mood models.

Requests that describe large sudden changes in emotion will be processed instantly. For other requests, the emotion state will gradually change into the requested state. Two outputs for the robot animator are generated: motion parameters and the current emotion. The motion parameters are generated in the emotion mixing part as they are directly related to the current emotion. For example, a cheerful robot has sharper and faster motions than a sleepy robot has. Due to time constraints the connection with the motion parameters has not been implemented in the current prototype.

2) Gaze Mixing: For the gaze mixing part of HMMM, implemented on the EyePi platform (Fig. B.2), two mixing types are used: single-modal and multimodal. The single-modal mixing processes multiple saliency maps that may come from various sources such as low level autonomous perceptual attention models, and high level deliberate attention in context of the dialogue. The autonomous perceptual attention models compute dynamic saliency over a continuous time frame, based on detected movement in a video feed. Deliberate attention maps are driven by dialogue actions at specific points in time. All maps are combined into a single map, keeping their original data intact, and fed into the mixer.

Due to small camera movements, but also movements of the object itself, the location of salient points will move over time. In order to track salient points in time as they move across the scene, we search for points that are within a certain threshold distance of existing salient points. In such cases, we update the center of the existing salient point, and update its weight to include the weight of the new point. This method prevents the ghosting of the different points when there is a moving body in front of the camera and it also functions as a smoothing function for the final EyePi movement.

To emulate more lifelike behaviour, we implemented a penalty and reward system for appropriately updating the weights of detected salient points. We distinguish between gradual and instantaneous penalties and rewards. The described behaviour is sketched in Fig. 4. Gradual penalties use a logarithmic function to decrease the weight of the most salient point, causing it to gradually become less salient over time. Instantaneous penalties and rewards are given when the focus of the agent's attention switches to a different point. This switching occurs when the weight of the old most salient point drops below (or is surpassed by) the weight of a new salient point. To prevent fast flipping between the two spots, the new most salient point receives an instantaneous reward and the old salient point receives an instantaneous penalty. The old point then starts to receive a gradual reward, until it is restored back to its original weight. Gradual rewards use a logarithmic function to increase the weight of old salient points.

The multimodal mixing of gaze works as follows. After selecting the most salient point, which is send onwards to be used as gaze target, this may also interact with the emotion models and the head gesture module. The autonomous generated map from the internal camera can induce 'shocked behaviour' by the robot, which leads to emotional response and a small expressive head movement. Finally, execution of gaze behaviour can be blocked in case specific gestures are active that would not be understandable when combined with gaze behaviour (see below).

3) Sequence Mixing: The final mixing part of HMMM, sequence mixing (Fig. B.3), handles both external requests and request from the gaze part. Sequences are pre-defined motions, which have specific motion definitions and requirements for every available actuator. Every robot platform will need its own definitions for all sequences in order to complete the mixing step. The definitions are specified on actuator level and they have one of the following classifications:

- *Required absolute motion* Absolute motion is required to complete the sequence. If it is not possible to do this, the sequence request must be rejected. It is impossible to mix this actuation with any other that controls the required actuator.
- *Not required absolute motion* This motion is still an absolute motion, but on conflicts it can be dropped.
- *Relative motion* As this motion is relative, it can be added to almost every other moved by adding its value. When an actuator is near its limit, the actuation can be declined.
- *Don´t care* The actuator is not used, so the sequence does not care about it.

Every sequence request has its own identifier, which is used in the feedback message in order to identify the feedback for the external software.

The first possible rejection is done based on these classifications. The current queue will be checked and the information from the requested sequence is retrieved from the database. If there are any conflicts in actuator usage that can not be solved, the request will be rejected. The second possible rejection is based on the timing of the requested sequence. If the timing can not be met, the request will be rejected. Both rejections will be send back to the requester using a feedback message.

If the sequence has passed both the actuator and timing check, the sequence planner will put it in the queue for execution. An acknowledgement request is send back to the requester and the processing of this specific request stops for the moment.

The second part of the sequence mixing is no longer directly part of the mixing process itself. There is a constantly running process which will activate sequences when they are allowed to start. When a sequence is started, feedback is send to the original requester that the sequence is started. The output to the animator contains both the sequence and possibly adjusted parameters in order to make the timing. The sequence is also transported to the gaze mixing part to operate the blocking behaviour there. The animator itself also sends feedback requests on animation strokes. Once a sequence is stopped, the sequence executor provides that action as feedback.

4) Animator: All mixing parts have one output in common: an output to an animator part. The animator is implementation specific and will differ per robot, but it needs to take the generated output from HMMM as input. These are the same as the input of the HMMM part, yet mixed, and they should not clash. An extra data channel is added: motion parameters to adjust speeds of the movements. Note that the animator also has an feedback output, required for progress feedback.

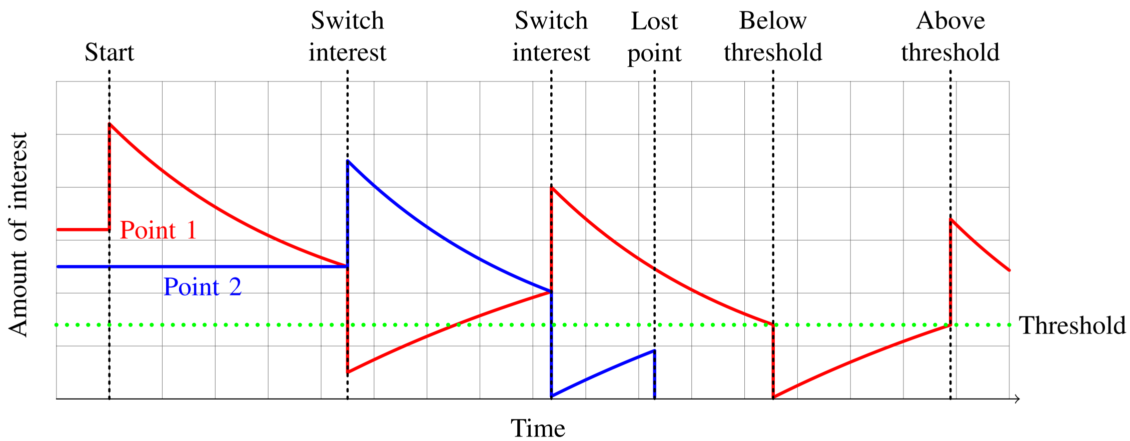


Fig. 4: A sketch of the interest for two interest points over time.

Figure B.4 provides a schematic overview of the animator as implemented for the EyePi.

VII. Discussion and Conclusion

In this project, we set out to mix deliberative and (semi)autonomous behaviour, in order to achieve fluent, multi-party, human-robot interaction. By extending the state-of-the-art BML realiser *AsapRealizer* and implementing the priority design pattern in the dialogue manager *Flipper*, we were able to achieve this. In the receptionist scenario, the robot showed fluent behaviour when assisting one interlocutor, and at some point during in the conversation switching to assist the bystander instead, when the robot recognised that the bystander was trying to attract its attention.

With our implementation, the role of the traditional role of the robot is transformed from a puppet, that always needs an puppeteer, into an actor which tries to follow its director. It interprets the requests and tries to execute them as best as possible. This way autonomous behaviour, such as breathing motions, eye blinking and (temporarily) gazing at interesting objects, will be combined but can also override the requests resulting in fluent and lifelike robot behaviour.

With the extension of *AsapRealizer* and the design pattern implementation in *Flipper* to handle interrupts during a conversation, a dialogue designer can now create responsive, lifelike and non-static dialogues, while only having to specify the deliberate behaviours.

This work was presented in the context of social robots. However, by virtue of the architecture of modern BML realisers, the approach will benefit also interaction with other embodied agents, such as Virtual Humans. To ensure that the *AsapRealizer* will also stay relevant for use with Virtual Humans, we have started working on the coupling with the Unity3D game engine[1] and editor which is a popular, state-of-the-art choice for virtual and mixed reality applications, both in research and industry. In the R3D3 project, the approach presented here will be used to govern the interaction between human users and a duo consisting of a robot and a virtual human. This project revolves around having such a duo take on receptionist and venue capabilities. HMMM will ensure that the envisioned interactions run smoothly and will be able to incorporate multiple users at the same time.

[1] http://unity3d.com/

Acknowledgements

The authors would like to thank the eNTERFACE '16 organisation, especially Dr. Khiet Truong, and the Design-Lab personnel. This publication was supported by the Dutch national program COMMIT, and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688835 (DE-ENIGMA), and the European Union Seventh Framework Programme (FP7-ICT-2013-10) under grand agreement No 611971 (EASEL).

Appendix A
Dialogues with the Receptionist Robot

Section III discussed the outline of the scenario we used to demonstrate our work on multi-modal mixing. In this Appendix, we explain the dialogues in more detail.

As explained in the subsection on the system architecture, the cascading triggering of the Flipper templates eventually leads to the dialogue templates being triggered (see section IV). Our method for realising the dialogue consists of two parts: firstly, dialogue management through *conversational phases*; secondly, behaviour planning through *behavioural intents*.

A. Dialogue Management

The dialogue with the robot revolves around users having a goal of visiting one of two doctors. The robot assists users in finding their way to their appointments. Fig. A.1 shows the setup of the interaction. Building on our ideas of a suitable scenario for HMMM, our scenario consists of several *conversational phases*, see Fig. A.2. Table A.1 lists the realisations of the robot's behaviour during each of its actions.

In the interaction with the robot, the first phase is the *Initialisation* phase, which is invisible to users. Here, when

Fig. A.1: Overview of the interaction setup: the Zeno R25 robot, the interlocutor and bystander, the Kinect, and two nameplates for the doctors to the sides of the robot.

the system is started, the initial world and user model are set in Flipper's information state. This happens internally, without any behaviour of the robot being shown. The *Welcome* phase consists of two actions: acknowledging and greeting the user. Code Listing C.1 shows the Flipper dialogue template which governs the behaviour of the robot (see Appendix C). To be triggered, it requires three preconditions to be met. Firstly, the current conversational phase which the current user or *interlocutor* is situated in must be *Welcome*. Additionally, the conversational substate must not yet exist, as this it has not been created at the start of the scenario. This template can only follow up on the first phase of the interaction and not during any following phases, during which this substate does exist. All of our dialogue templates use this construction to order the steps in the interaction. Thirdly, the distance of the current interlocutor to the robot is checked. When the system has been started, the *SceneAnalyzer* continuously scans the scene and updates the world model in the information state. When an interlocutor is detected, she gets a unique ID and she is tracked in the scene. We defined several zones of proximity based on Hall's interpersonal distances [31], with the outer boundary of social space being 3.7 meters and that for personal space being 2.1 meters from the robot.[2] The SceneAnalyzer determines the distance of the user to the robot. When the user comes closer than 3.7 meters, a Flipper template triggers which sets the user's interpersonal distance to social.

Together, these three preconditions trigger a number of effects. As described above, the conversational phase (and substate) are updated. To handle multiple users, the priority of this action is set to a particular number. This is explained in Section VI-B. Finally, a behavioural intent is added to a queue of actions to be carried out by the robot. We explain this functionality in the following subsection. When the user steps into the personal distance of the robot, the next template triggers, namely the one causing the robot to greet the user.

The remaining conversational phases follow a similar structure. The robot's goal during the *Instruct* phase is to indicate what users should do in order to reach their appointment. Having welcomed a user, the robot instructs her to point to the nameplate of the doctor with whom she has an appointment (the `Instruct` template). The robot synchronises verbal and non-verbal behaviour to both point at and gaze at each of the nameplates in turn. The SceneAnalyzer detects whether one of the user's hands points either left or right. This information is further processed and when the user has made a choice, the next phase can be triggered. If this is not the case, the robot waits a certain amount of time (20 seconds) before re-iterating its instructions (`InstructAgain`). Again, if the user makes a choice, the dialogue progresses to the next phase. If she fails to express her choice within a certain amount of time (20 seconds), the robot apologises for not being able to help her out and directs her to a nearby human to further assist her (`DismissAfterInstruct`). Then, the robot idly waits until the user leaves and a new user enters.

When the user has indicated her choice, she enters the *Direct* phase, receiving directions on how to get to her appointment (`Direct`). Based on the user's choice, the robot utters a sentence and gazes and points in the direction in which the user should head. Similar to the previous phase, the robot either repeats its directions (`DirectAgain`) or redirects the user to someone else (`DismissAfterDirect`) when the user fails to move in the correct direction after a set amount of time (20 seconds). Again, the SceneAnalyzer is responsible for detecting the user's behaviour. If, after being directed or being directed again, the user walks into the wrong direction, the robot will call her back to it (`DirectAfterIncorrectDirection`). This happens when she leaves the 'personal' space of the robot and exits the interaction range in the opposite direction of which she

[2]These zones correspond to Hall's far phase and close phase in social distance, respectively [31]; in our setup, we renamed them for clarity.

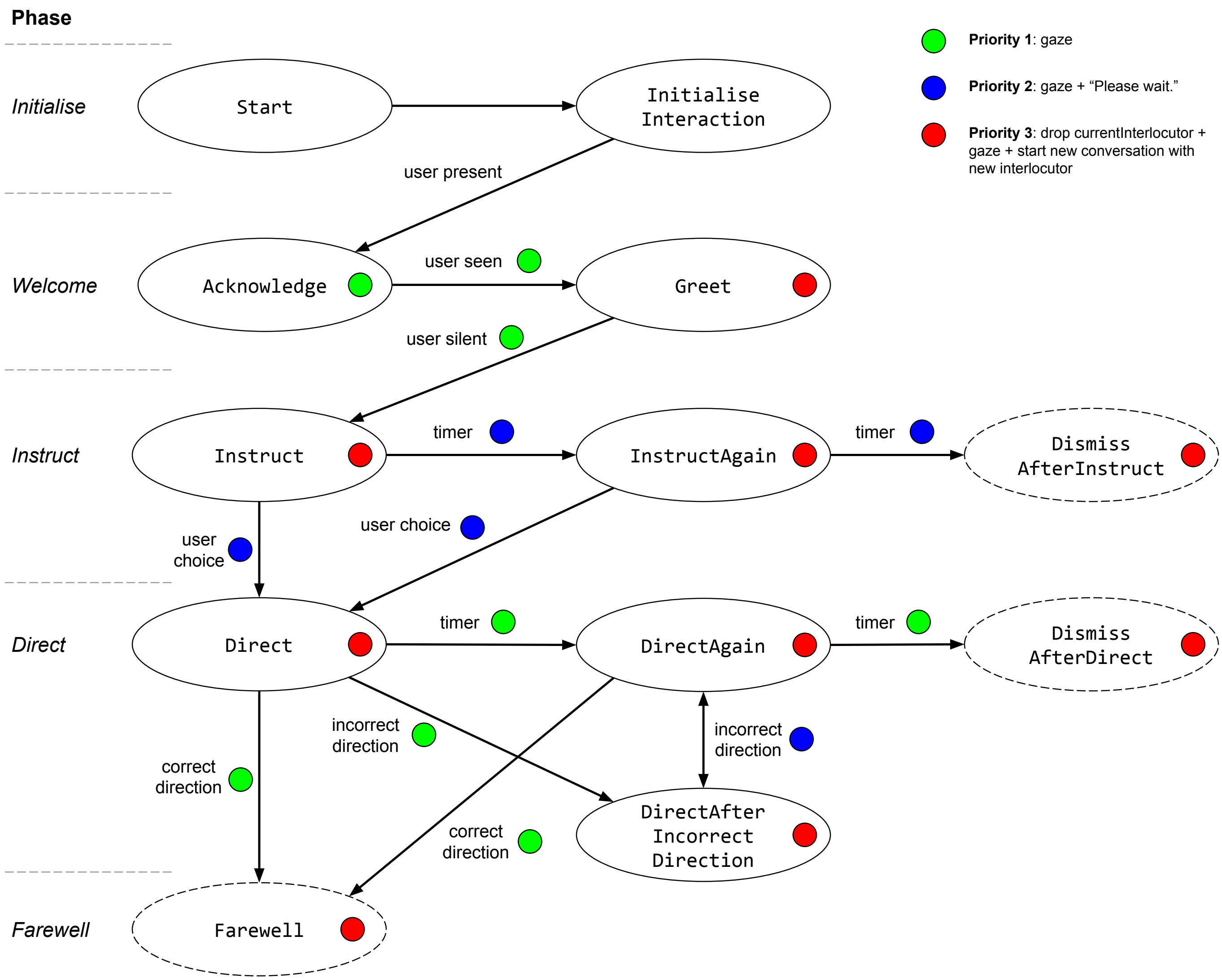


Fig. A.2: Schematic overview of the receptionist scenario.

should be headed to. Instead, if the user leaves the robot's personal space in the correct direction, the robot utters a friendly goodbye and waves her off (`Farewell`).

B. Behaviour Planning

When Flipper dialogue templates trigger, their effects are executed. As described in the previous subsection, behaviours of the robot are triggered through *behavioural intents*, see Code Listing C.1. Previously, Flipper used `behaviour`-tags for the specification of BML behaviour in these templates. We replaced these tags with behavioural intents to accommodate for different realisations of behaviour, e.g., by different robots or virtual agents. To this end, the intents are of a higher order specification than the explicit BML commands. In Code Listing C.1 (see Appendix C), the robot's intent is to acknowledge the current interlocutor. A request with this intent is added to a queue of behaviours to be planned by AsapRealizer. Then, it is up the realiser to plan this behaviour for a specific robot or virtual agent. The advantage of this approach is that the dialogue remains realiser-agnostic: for a different entity, the BML needs to be specified for each behaviour, separated from the dialogue templates. This planner uses Flipper templates to take the first intent in the queue of planned intents and checks which type of behaviour should be planned. Based on this information, it carries out optional translations of information from the SceneAnalyzer. In the case of the `Acknowledge` intent, the coordinate system of the SceneAnalyzer data is translated to the coordinate system of the Zeno robot, so that it is able to look at the correct position of the user's head. Then, this template calls AsapRealizer to realise this behaviour using BML. Code Listing C.2 (see Appendix C) shows the BML behaviours used by the Zeno and the EyePi robots, respectively, for the `Acknowledge` intent.

Appendix B
Heterogeneous Multilevel Multimodal Mixing

Figures B.1 (emotion), B.2 (gaze) and B.3 (sequence) contain the schematic overviews of HMMM. The three parts handle different behaviour requests and mix them into viable

TABLE A.1: The Realisations of the Behaviour of the Zeno Robot, for Each of the Intents Mentioned Shown in Fig. A.2.

Intent	Behaviour realisation	
	Verbal	**Non-verbal**
`Acknowledge`	(None.)	Look at user.
`Greet`	Hello, my name is Zeno.	Wave at user.
`Instruct`	Please point at the sign with your doctor's name on it. Either this sign (1) on your left, or this sign (2) on your right.	At point (1), look at the sign on the left and point at it with left arm; at (2), look at the right sign with the right arm and point at it.
`InstructAgain`	My apologies, maybe I was not clear. Please point at the sign with your doctor's name on it. Either this sign (1) on your left, or this sign (2) on your right.	At point (1), look at the sign on the left and point at it with left arm; at (2), look at the right sign with the right arm and point at it.
`DismissAfterInstruct`	(1) I'm sorry, I'm not able to help you out. My capabilities are still limited, so I was not able to understand you. (2) Please find a nearby human for further assistance.	At point (1), make a sad face; at (2), make a neutral face.
`Direct`	Please go to the left/right (a) for doctor Vanessa/Dirk (b).	Depending on the user's choice, instruct the user, and look and point at the corresponding direction the user should head in (a, b).
`DirectAgain`	(1) My apologies. Maybe I was unclear. (2) Please go to the left/right (a) for doctor Vanessa/Dirk (b).	At point (1), make a sad face; at (2), make a neutral face. Depending on the user's choice, instruct the user, and look and point at the corresponding direction the user should head in (a, b).
`DismissAfterDirect`	(1) I'm sorry, I'm not able to help you out. Please find a nearby human for further assistance. (2)	At point (1), make a sad face; at (2), make a neutral face.
`DirectAfterIncorrectDirection`	(1) Sorry, but you're headed the wrong way! Please come back here. (2)	At point (1), make a confused face; at (2), make a neutral face.
`Farewell`	(1) That's the way! (2) Goodbye!	At point (1), make a happy face; at (2), wave at the user.

robot behaviour, while still using autonomous behaviour, as discussed in Section VI-C.

APPENDIX C
CODE LISTINGS

This appendix contains code snippets from the project. Code Listing C.1 shows the dialogue template for the acknowledgement behaviour of the robot. Code Listing C.2 shows the BML for the acknowledge behaviour for both the Zeno and eyePi robots.

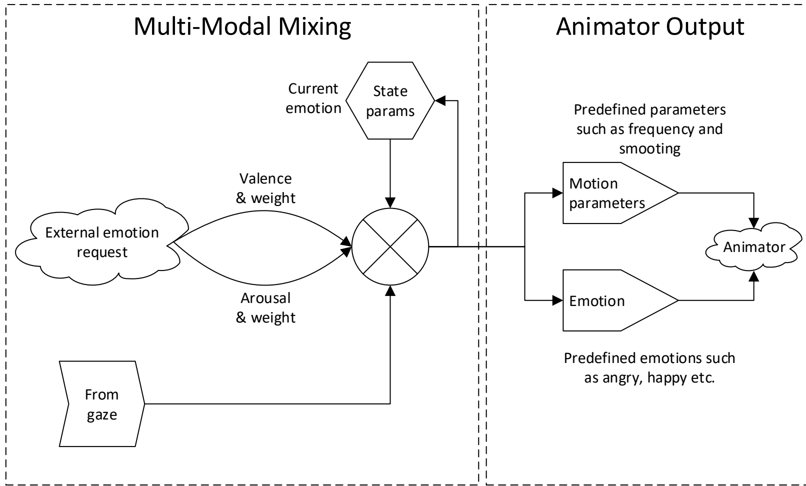


Fig. B.1: Schematic overview of the emotion mixing.

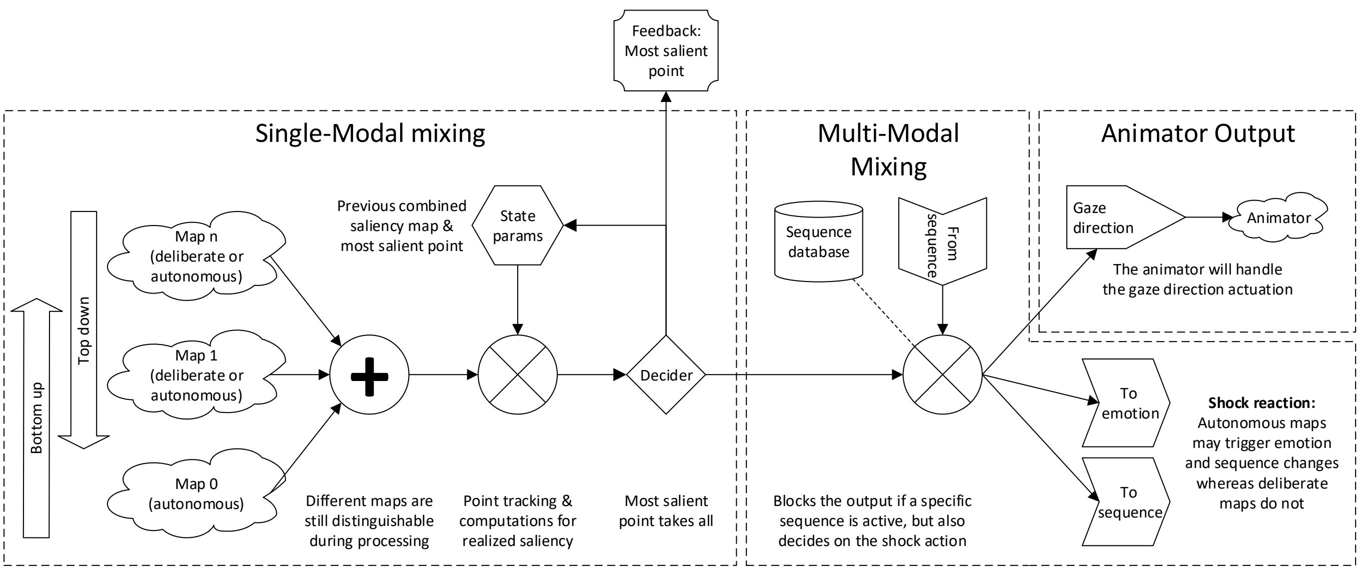


Fig. B.2: Schematic overview of the gaze mixing.

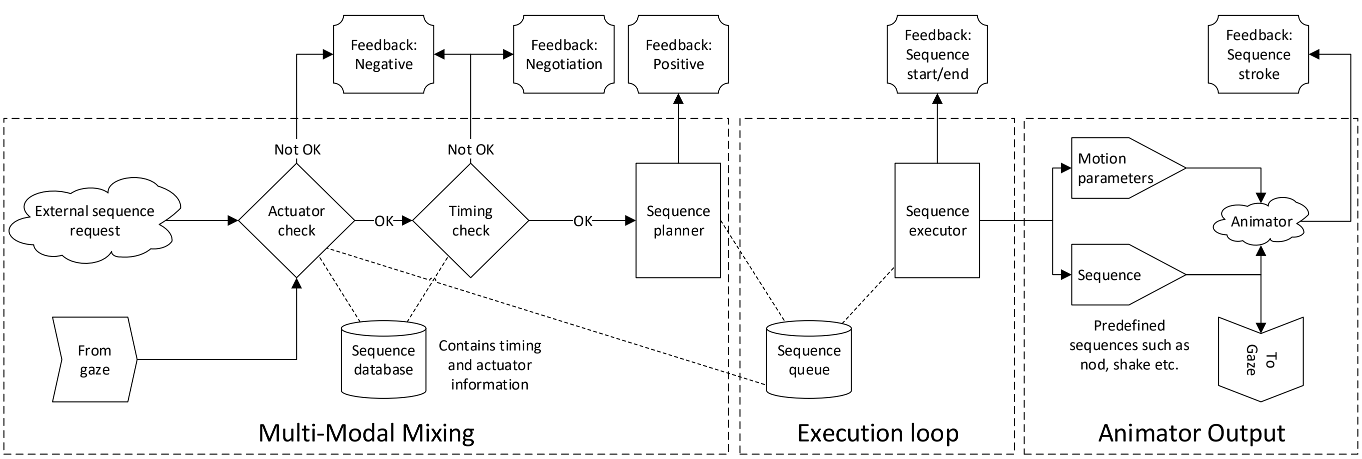


Fig. B.3: Schematic overview of the sequence mixing.

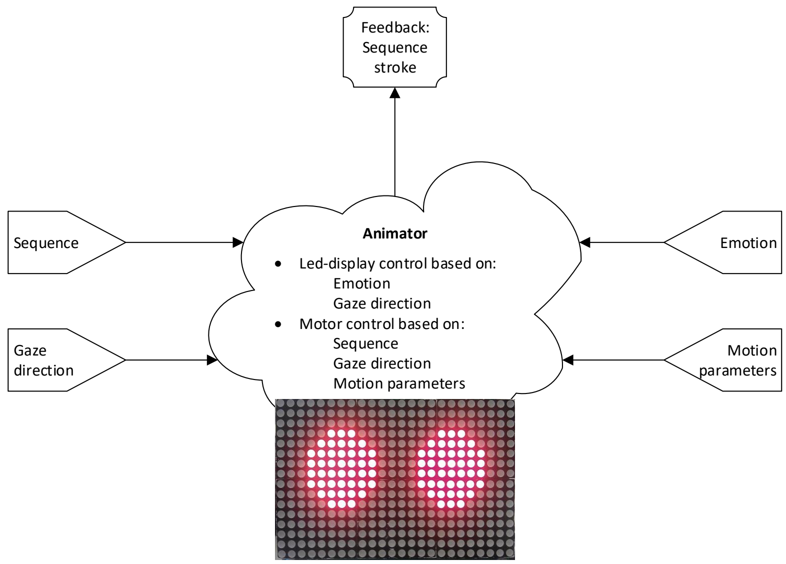


Fig. B.4: Schematic overview of the animator.

Listing C.1: The Flipper Dialogue Template for the Acknowledgement Behaviour of the Robot.

```
<!-- When a user has been detected, but is not yet within interaction range,
      let the robot look at the user. -->
<template id="hmmmAcknowledge" name="hmmmAcknowledge">
  <!-- These preconditions must be satisfied before the template triggers. -->
  <preconditions>
    <compare value1="$interactionContext.currentInterlocutor.cstate" value2="welcome"/>
    <compare value1="$interactionContext.currentInterlocutor.csubstate" comparator="not_exists"/>
    <compare value1="$interactionContext.currentInterlocutor.socialDistance" value2="SOCIAL"/>
  </preconditions>

  <!-- These effects result from the template triggering. -->
  <effects>
    <!-- Update the conversational substate so the template will not trigger again. -->
    <update name="$interactionContext.currentInterlocutor.csubstate" value="acknowledged"/>
    <update name="$interactionContext.currentInterlocutor.interactionStarted" value="TRUE"/>

    <!-- The priority of this action is set. -->
    <update name="$conversationalContext.priority" value="1"/>

    <!-- The intent of the acknowledgement behaviour is added to a queue of planned behaviours.. -->
    <update name="$isTemp.newRequest.r.intent" value="acknowledgeInterlocutor"/>
    <update name="$isTemp.newRequest.r.target" value="currentInterlocutor"/>
    <update name="$isBehaviourPlanner.requests._addlast" value="$isTemp.newRequest.r"/>
    <remove name="$isTemp.newRequest.r"/>
  </effects>
</template>
```

Listing C.2: BML Behaviours for Realisation of the Acknowledge Intent by the Zeno Robot and by the eyePi.

```
<!-- BML for Zeno -->
<bml id="$id$" xmlns="http://www.bml-initiative.org/bml/bml-1.0"
    xmlns:sze="http://hmi.ewi.utwente.nl/zenoengine">
  <!-- Zeno will look at the user's head position. -->
  <sze:lookAt id="lookAtCurrentInterlocutor"
    x="$interactionContext.currentInterlocutor.x$"
    y="$interactionContext.currentInterlocutor.y$"
    start="0" end="0.2"/>

  <!-- After having looked at the user for two seconds, Zeno will look to the front again. -->
  <sze:lookAt id="lookToTheFront" x="0.5" y="0.5" start="2" end="2.2"/>
</bml>

<!-- BML for eyePi -->
<bml id="$id$" xmlns="http://www.bml-initiative.org/bml/bml-1.0"
    xmlns:epe="http://hmi.ewi.utwente.nl/eyepiengine">
  <!-- eyePi will look at the user's position. -->
  <epe:eyePiGaze id="lookateyepi" x="$x$" y="$y$" start="0" end="0.1"/>
</bml>
```

References

[1] B. van de Vijver, "A human robot interaction toolkit with heterogeneous multilevel multimodal mixing," Master's thesis, University of Twente, the Netherlands, 2016. [Online]. Available: http://purl.utwente.nl/essays/71171

[2] V. Charisi, D. P. Davison, F. Wijnen, J. van der Meij, D. Reidsma, T. Prescott, W. Joolingen, and V. Evers, in *Proceedings of the Fourth International Symposium on "New Frontiers in Human-Robot Interaction"*, 2015, pp. 331–336.

[3] H. van Welbergen, D. Reidsma, and S. Kopp, "An incremental multimodal realizer for behavior co-articulation and coordination," in *Proceedings of the 12th International Conference on Intelligent Virtual Agents*, 2012, pp. 175–188.

[4] G. Hoffman, "Ensemble: Fluency and embodiment for robots acting with humans," Ph.D. dissertation, Massachusetts Institute of Technology, 2007.

[5] G. Hoffman and C. Breazeal, "Effects of anticipatory perceptual simulation on practiced human-robot tasks," *Autonomous Robots*, vol. 28, no. 4, pp. 403–423, 2009.

[6] G. Klein, P. J. Feltovich, J. M. Bradshaw, and D. D. Woods, "Common ground and coordination in joint activity," *Organizational simulation*, vol. 53, pp. 139–184, 2005.

[7] S. Kopp, "Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors," *Speech Communication*, vol. 52, no. 6, pp. 587–597, 2010.

[8] D. Heylen, "Head gestures, gaze, and the principles of conversational structure," *International Journal of Humanoid Robotics*, vol. 3, no. 3, pp. 241–267, 2006.

[9] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, 2006, pp. 205–217.

[10] D. Reidsma and H. van Welbergen, "AsapRealizer in practice — A modular and extensible architecture for a BML Realizer," *Entertainment Computing*, vol. 4, no. 3, pp. 157–169, 2013.

[11] H. van Welbergen, D. Reidsma, and J. Zwiers, "Multimodal plan representation for adaptable BML scheduling," in *Autonomous Agents and Multi-Agent Systems*, 2013, pp. 305–327.

[12] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[13] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, no. P1, pp. 22–35, 2015.

[14] L. Riek, "Wizard of Oz studies in HRI: A systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, 2012.

[15] D. Bohus and E. Horvitz, "Dialog in the open world," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*, 2009, pp. 31–38.

[16] ——, "Multiparty turn taking in situated dialog: Study, lessons, and directions," in *Proceedings of the SIGDIAL 2011 Conference*, 2011, pp. 98–109.

[17] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, no. 1, pp. 1–26, 1992.

[18] A. Csapo, E. Gilmartin, J. Grizou, J. Han, R. Meena, D. Anastasiou, K. Jokinen, and G. Wilcock, "Multimodal conversational interaction with a humanoid robot," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 667–672.

[19] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 2009, pp. 61–68.

[20] E. Goffman, "Footing," *Semiotica*, vol. 25, no. 1–2, pp. 1–30, 1979.

[21] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, pp. 1–33, 2012.

[22] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka, "Precision timing in human-robot interaction: coordination of head movement and utterance," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 131–140.

[23] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating vision and audition within a cognitive architecture to track conversations," in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 2008, pp. 201–208.

[24] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 40, pp. 696–735, 1974.

[25] G. Skantze, "Predicting and regulating participation equality in human-robot conversations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 196–204.

[26] V. Vouloutsi, M. Blancas, R. Zucca, P. Omedas, D. Reidsma, D. Davison, V. Charisi, F. Wijnen, J. van der Meij, V. Evers, D. Cameron, S. Fernando, R. Moore, T. Prescott, D. Mazzei, M. Pieroni, L. Cominelli, R. Garofalo, D. De Rossi, and P. F. M. J. Verschure, "Towards a synthetic tutor assistant: The EASEL project and its architecture," in *Conference on Biomimetic and Biohybrid Systems*, F. N. Lepora, A. Mura, M. Mangan, F. P. Verschure, M. Desmulliez, and J. T. Prescott, Eds., 2016, pp. 353–364.

[27] A. Zaraki, D. Mazzei, N. Lazzeri, M. Pieroni, and D. De Rossi, "Preliminary implementation of context-aware attention system for humanoid robots," in *Conference on Biomimetic and Biohybrid Systems*, 2013, pp. 457–459.

[28] M. ter Maat and D. Heylen, "Flipper: An information state component for spoken dialogue systems," in *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, 2011, pp. 470–472.

[29] H. van Welbergen, D. Reidsma, Z. Ruttkay, and J. Zwiers, "Elckerlyc," *Journal of Multimodal User Interfaces*, vol. 3, no. 4, pp. 271–284, 2009.

[30] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[31] E. T. Hall, *The hidden dimension*. New York, NY: Bantam Doubleday Dell Publishing Group, 1966.

Design and Development of a Physical and a Virtual Embodied Conversational Agent for Social Support of Older Adults

Marieke M.M. Peeters; Vivian Genaro Motti; Helena Frijns; Siddharth Mehrotra; Tuğçe Akkoç; Sena Büşra Yengeç; Oğuz Çalık; and Mark A. Neerincx

***Abstract*— Populations in developed societies show an increasingly higher life expectancy across the globe. To support older adults to live longer and healthier lives in the familiar surroundings of their homes, technological developments, such as robots and avatars, have a great potential.**

To investigate long-term interactions between older adults and a "bi-bodied conversational agent" (an agent that has both an avatar and a robot embodiment), a user-centred design approach was employed in the design and development of a conversational agent. Firstly, the requirements of the agent were elicited through a set of focus groups with the target users – older adults.

Then, the agent was iteratively designed and implemented: a robot body and avatar body were created. Finally, a wizard-of-Oz control panel was created to control and compare each of the two bodies. Current research outcomes describe the elicited requirements baseline of a bi-bodied conversational agent for older adults. Future research involves the use of this set-up to investigate long-term interaction between older adults and a bi-bodied conversational agent.

***Index Terms*— Robot; Social Agents; Avatar; Wizard of Oz; Conversational Interfaces; Older Adults; Human Computer Interaction; User Centred Design; Embodied Conversational Agents**

I. Introduction

OVER the next two decades, almost 20% of the total population in developed countries will be of age 65 and above. As a result of the rapidly growing population of older adults and lack of caregiving resources, governments and healthcare providers are eager to find new ways of providing high quality elderly care, while reducing the associated costs.

One promising direction in finding a solution to the problem of quality care for an increasingly aging population is the use of innovative technology. Robots and avatars are good examples of technological solutions with a large potential in this field. By integrating sensing, mechatronics, and communication technologies, they are able to support older adults during their daily activities, facilitating connectedness between older adults and their social environment.

Despite the large potential that technology has for the aging population, previous research has shown that older adults are usually slower than their younger counterparts to adopt novel technologies [1,2]. This hesitation, potentially caused by inaccessible or user-unfriendly solutions, excludes a large part of older adults from benefiting from technology. In fact, regardless of their purposes –medical, social, or household– , to assure higher adoption and acceptance rates, future technological solutions for older adults need to carefully consider the needs, requirements, and capabilities of the elderly generations [3,4,5].

One of the design challenges in technology for older adults is having a unified definition for the target group. In the literature, the age used as a cut-off-point for elderly people ranges from fifty [6] to eighty-eight [7]. This wide age range challenges researchers to address diverse needs of older adults who are among the largest but most under-represented user-groups in scientific studies.

Previous research indicates that embodied conversational agents (ECAs) (e.g. robots or avatars) are especially promising in supporting elderly to age in their home environment [8,9,10,11,12]. ECAs can offer older users a natural and intuitive interface to interact with complex technology –such as smart homes, web applications, and so on– thereby facilitating interaction between the elderly and the technology [13]. An ECA can adopt many forms and embodiments, yet for the current purpose of designing an intuitive interface for older adults, two formats stand out:

1) a virtual body in the form of an avatar
2) a physical body in the form of a robot

The main difference between these two options is that a robot is physically present in 3D space, as a tangible interface, whereas a virtual avatar is only present on a screen. As a result, a robot and an avatar also have different affordances: an

This paper was submitted on September 12th,2016. The work described in this paper was funded by the 4TU.Humans&Technology research center.

Marieke M.M. Peeters is a research scientist Artificial Intelligence at TNO in Soesterberg, the Netherlands (marieke.peeters@tno.nl).

Vivian Genaro Motti is an assistant professor at George Mason University in Fairfax, United States of America (vmotti@gmu.edu).

Helena Frijns is an MSc student at Leiden University in Leiden, the Netherlands (h.a.frijns@umail.leidenuniv.nl).

Siddharth Mehrotra is a postgraduate student at the International Institute of Information Technology in Hyderabad, India (the4daspect@hotmail.com).

Tuğçe Akkoç is an MA student at Boğaziçi University in Istanbul, Turkey (akkoctugce@gmail.com).

Sena Büşra Yengeç is an electrical engineer in Ankara, Turkey (senabyengec@gmail.com).

Oğuz Çalik is an electrical engineer in Ankara, Turkey (o_calik@outlook.com).

Mark A. Neerincx is a full professor at Delft University of Technology in Delft, the Netherlands. He is also a principal scientist at TNO in Soesterberg, The Netherlands (mark.neerincx@tno.nl).

avatar is able to provide help in a virtual environment, whereas a robot is able to provide help in a physical environment. Yet, in addition to the inherent differences in their capabilities, having a physical body may also affect the user-agent relationship in unforeseen ways. Previous research suggests that, compared to the avatar interaction, the interaction with a robot is more engaging [14]. Furthermore, a robot is often perceived as more helpful and enjoyable, as something with a social presence that exists and is experienced as real [15,16,17]. Research with university students found that people had more positive interactions and engaged more with relaxation exercises when these were presented by a robot with an embedded screen, as opposed to the exercises being presented on a tablet [18]. Yet despite the seemingly positive effects of robotic ECAs, most of the available systems developed for older adults still involve screen-based interfaces instead of robots [19].

The appearance of social agents – be they avatars or robots – is often human-like or animal-like, and it can be challenging to find the right balance when trying to offer a familiar and intuitive communicative interface without raising too high expectations or becoming unsettling [20,21].

An interesting option to get the best of both worlds, is to create a conversational agent that can take control of two bodies, one of which is a virtual avatar body that can help a user in the virtual environment, and the other a robotic body that can aid the user in his/her physical environment. Older adults prefer to interact with ECAs through more natural and familiar modalities [22]. The usability of both bodies may vary according to the needs of the environment, but the user's perception of the agent can be established by having similar characteristics and functions in different bodies. It is supported that users perceive such an agent as a continuous identity even if it migrates between different embodiments [23].

Despite a large body of research on social robots, and a consensus on their large and promising potential to help older adults to age in place, there is a limited understanding on the perspectives of older adults about these technological solutions, especially concerning the benefits and drawbacks related to ECAs in the form of a virtual agent or avatar versus ECAs in the form of robots. Therefore, the *research objective* of the work described in this paper is to design and develop a research set-up that enables researchers to further investigate long-term interactions between elderly people and ECAs with both a physical and a virtual body. Having a physical and a virtual solution of a conversational agent permits researchers to conduct user studies with older adults to assess multiple quality criteria, including the acceptance and usefulness of the solutions, as well as the costs and benefits associated to each form.

This paper is structured as follows. Section II describes related work on the design of avatar and robotic embodiments for ECAs. Section III introduces the research approach. Section IV presents the design of the agents as envisioned by the researchers. Section V describes the method employed to elicit a list of user requirements that helped shape our design of the ECA. Section VI presents the outcomes of the research. And Section VII reflects on the work described in this paper, by discussing the implications of the findings, plans for future research, and concluding remarks.

The key contributions of this research include: (1) a theoretical contribution (design decisions and user study results), (2) a computational solution (virtual agent), and (3) a physical prototype (robot).

II. Related Work

Previous research has explored and compared various available ECA-based solutions for elderly users in terms of, e.g., feasibility of implementation, usefulness, and user acceptance. A review of the literature on ECAs for elderly reveals the following state of the art in terms of human factors and affordances that are of relevance to the design of ECAs.

Bennet & Šabanović (2014) investigated human-like robotic faces [24]. Their findings suggest that eyebrow, eye and mouth movement are sufficient for recognition of emotions. Additional neck movements may further increase correct recognition of expressed emotions. A robot design might not be acceptable if users perceive it as "childish" [25]. Volonte et al (2016) investigated a cartoon-shaded versus a sketch-like versus a more realistically shaded virtual avatar. They found that the rendering of the avatar has an effect on measured affect and the way participants perceived the personality of the avatar [26]. Paauwe et al. (2015) suggest that perceived affordances (i.e. realism in terms of participants being able to imagine using the robot in the real world) are more important than modelled realism (in terms of visual similarity to existing "agents" like nurses) to the usability of a robot [27]. Wu et al. (2014) conducted focus groups and semi-structured interviews with elderly with mild cognitive impairment to explore their attitudes toward assistive robots. They found that many of the participants were hesitant to the proposed technology and its functionalities, arguing that similar functionality was already available on other devices. The participants could not imagine using robots themselves, but could see some use for elderly who were more impaired than they were. Wu et al. suggest that in order to maintain elderly people's autonomy, it is better to have robots fulfil a supplementary function, rather than replace human functions [28].

A recent EU-financed (h2020) research project called PAL4U[1] uses a robot as well as an avatar of the robot to motivate children with diabetes to better manage their health [29]. PAL uses the virtual robot to allow for the children to communicate and interact with the robot through their smart devices, resulting in lower costs. When the children come back to the hospital for checks and follow-ups, the real robot is there using a user model that is shared with the virtual robot, to ensure consistency in its behaviour toward the child. The PAL project illustrates how two bodies can be complementarily used in different settings and for different purposes: at home the virtual assistant helps children to fill out their diary, while at the hospital the physical robot plays games with the children.

[1] http://www.pal4u.eu/

III. APPROACH

To ensure that the users' perspectives are considered from an early design stage onwards, this study followed a user-centred design approach. Four design phases were combined: 1) conceptual design (framework and requirements), 2) brainstorming, 3) prototyping, and 4) user studies (questionnaires and focus groups).

The methodology for our design study, is inspired by the situated Cognitive Engineering (sCE) method [30,31]. The sCE method aims to support the design of intelligent human-computer interaction. It focuses on the iterative specification and refinement of a design rationale: a comprehensive argumentation underpinning the design choices made throughout the design process. A design rationale is situated in the task domain (operational demands), founded in theories from human factors, implemented in a technological solution, and evaluated through empirical evaluations.

The design process took place through a number of rapid iterative cycles, each cycle taking the outcomes of the previous cycle to refine the design ideas and test the outcomes with the end users. In the first stage, various possible designs were created by the research team based on an initial functional framework. In the stages that followed, the design ideas were incrementally refined and evaluated with the target users through a questionnaire and four focus groups. The stages are presented and discussed in detail in the next sections. Based on the end users' perspectives gathered throughout the design process, a semi-functional high-fidelity prototype of the ECA was built. This study received approval from the Human Research Ethics Committee from Delft University of Technology.

IV. DESIGN PREPARATION

The current study focuses on the design of a bi-bodied conversational agent for older adults. First, a functional framework (Figure 2) was collaboratively defined by the research team. This framework describes the basic components and functionalities of the to-be-developed set-up. As input controls, the user's behaviours are detected through sensors, e.g. a microphone (for audio and speech analysis), and proximity sensors (for motion sensing). Seeking to assess the feasibility of the two prototype formats, as well as their benefits and drawbacks, the following design requirements were elicited.

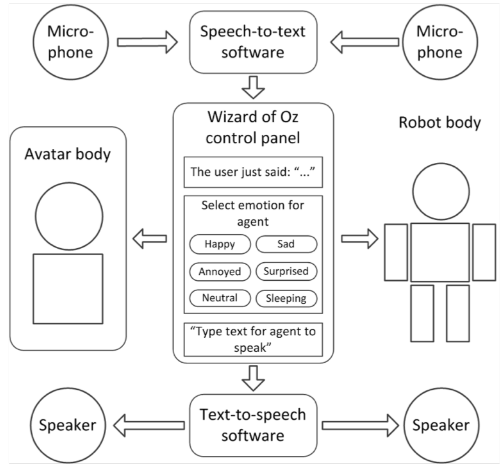


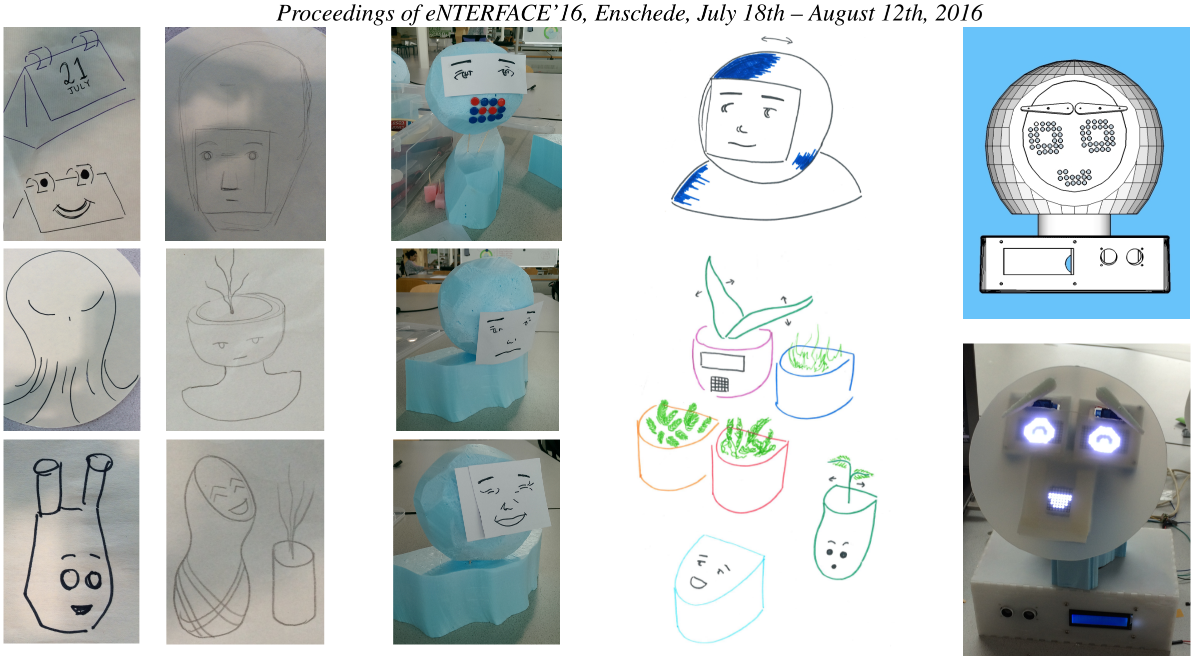
Fig. 2. Functional framework of the envisioned application enabling a comparative analysis of an embodied conversational agent for older adults with two embodiments (a virtual and a physical body)

A set of requirements, including features available and quality factors, were discussed in the early development phases. The functional requirements for the solution include:

- The robot body and avatar body should enable control through a Wizard-of-Oz (WoOz) [32] technique (i.e. unbeknownst to the user/participant, a human researcher mimics the envisioned interaction with the user before actually implementing the software or artificial intelligence needed to automatically produce such behaviour) ensuring a consistent behaviour for both embodiments:
    - The WoOz should be able to instruct the robot/avatar to talk
    - The WoOz should be able to change the facial expression of the robot/avatar
    - The WoOz should receive information about the conversation, e.g. facial expression of the user, contents of the conversation

The non-functional requirements for the solution include:

- The two prototype formats (i.e. the avatar body and the robot body) should be consistent in a way that users can identify the similarities between the two embodiments
- The solutions, including the agent bodies, voice, and facial expressions should be acceptable and likeable by elderly users.



(A) (B) (C) (D)

Fig. 1. Various designs and prototypes: (A) Six low-fidelity paper-based designs, ranging from objects, animals, abstract, human-like, and plant-like shapes (B) Three foam-based prototypes illustrating 3D sculptures of potential body formats (C) Pre-final designs for discussion in focus groups, human-like and plant-like designs, (D) 3D model of the final design (top) and final prototype (bottom)

Based on the initial framework, a brainstorm session was conducted by the research team, to propose and discuss the advantages and drawbacks of various alternative looks for the ECA design. The research team was composed of people with different backgrounds: computer science, human-centred design, electrical engineering, cognitive science, psychology, mathematics, and fine arts.

The brainstorm session was followed by a design session, in which 2-D paper-based prototypes (Figure 1 A – calendar-like, human-like, animal-like, plant-like, abstract and object formats) and 3-D foam-based prototypes (Figure 1 B – robot head and body compositions) were built to illustrate different ECA formats. Sketches of a human-like and a plant-like ECA were generated and discussed in the early design stages of the study (Figure 1 C).

Six paper-prototypes (Figure 1A) were sketched and discussed during a brainstorm session (the look ranged from human-like, to animals, plants, and abstract shapes). During the session the research team analysed the advantages and disadvantages (e.g. mobility, placement in the home, acceptability, aesthetics, stigma, associations, etc.) of each design regarding potential familiarity, comfort, costs, meaning, and ease of use. The researchers then selected two designs to be discussed in focus groups with the target users: a human-like design (Figure 1C above) and also a plant-like design (Figure 1C below).

V. USER STUDIES

A. Method

The design decisions for the solutions proposed were guided by four complementary design aspects, namely:

1) The appearance of the bodies, including the aesthetics, dimensions, textures, colours, and visual features. Decisions regarding these aspects were made through thorough discussions of the prototypes among the research team and potential users;
2) The functionality of the bodies, i.e. suitable and/or useful purposes of the ECA for older adults, in terms of features and functionalities. To decide on the possibilities, a set of user studies was conducted (combining questionnaires and focus groups);
3) The creation of the two bodies (physical and virtual one), decided after a cycle of proposal, test, development, and preliminary evaluation of the robot and the avatar; firstly discussing the feasibility of the construction (hardware and software aspects), given the resources' constraints of time, design, and budget, then focusing on implementation and testing;
4) The software for the WoOz panel: logical aspects (algorithm), technical implementation, as well as the integration, compatibility, languages (Dutch and English), and scalability aspects. All thoroughly discussed by the design team, resulting in a combination of existing tools (e.g. Pandora bot and MyRobotLab) and customized software solutions.

1) Appearance

Prototypes made out of paper and foam were created to explore different shapes, dimensions and facial expressions for the robot. Figure 1A and 1B illustrate alternative face and bodies for the robot as considered in the early stages of the project. The design of the appearance was then further refined through three research steps: a brainstorm session with the team, a questionnaire, and four focus groups. Each step is further explained in the following subsections.

a) Questionnaire

To gather general insights about the older adults' perspectives on ECAs (appearance, potential functionality, and previous experience), we conducted a preliminary online survey with 22 participants (11 men and 11 female, M_{age} = 67 years, SD_{age} = 2.5 years) from the Netherlands, Turkey and India). The questionnaire was distributed among family, friends, friends of friends, and through social media. The questionnaire provided a brief explanation of what an ECA is and how it might be of use to support in daily chores around the house. People were then asked whether they have any chores around the home that they could use some help with. Participants responded by mentioning various daily routines they engage in, and how they use calendars, mobile apps, sticky notes, and the placement of reminder objects around the home. People were then shown four of the designs created by the team of researchers, i.e. the humanoid, the animal-like, the plant-like, and the abstract design (also see Figure 1A). On asking about the prototype of the robot participants' opinions varied on what kind it should be: 31.8% preferred it to be humanoid, 45.5% for animal look, 13.6% for plant look and the remainder, 31.8% preferred it to be abstract.

b) Focus Groups

To gather specific feedback from older adults concerning existing ECAs and potential design formats, we also conducted four focus groups. As mentioned in the introduction, we acknowledge that the user group "elderly people" is a large and highly diverse group of people. For this reason we included people from various ages (60-85), men and women, and people who are still active and healthy as well as people with somatic problems. In total, 23 older adults participated in the four sessions (14 men and 9 females, M_{age} = 71 years, SD = 2.5 years) from the Netherlands and Germany. The participants were recruited through their garden club, via word-of-mouth, and through a healthcare organization in the Netherlands. Participants signed the informed consents before the sessions started. All focus groups consisted of different participants. Most of the participants in the first two sessions were highly educated (university level) and did not have any important cognitive or motor impairment, although some participants had hearing aids and visual impairment natural to their age. In the third and fourth focus group, participants were not highly educated and did show more serious somatic impairments such as multiple sclerosis.

The first two focus groups mainly focused at assessing the acceptability of the two paper prototypes (i.e. humanoid vs. plant/object – Figure 1C), however, before assessing the two designs with the participants, the researchers provided an introduction to the topic, discussing what constitutes a robot or avatar, and what purposes they might serve. After all participants had a clear idea of what a robot or avatar might be able to do for them, the two designs were shown and evaluated with the target group.

The first prototype showed a humanoid face on a screen that was embedded in a sphere (named astronaut-like design). This sphere represented the robot's head attached to a neck to connect the head with the robot's body. Participants were informed that the facial expression of the robot could change depending on its programmed mood. The second prototype showed robots that were plant-like or object-like (vase-like). Participants expressed a preference for the first prototype and immediately rejected the robotic functions in unanimated objects like vases or plants as those would infantilize them. Most participants preferred the human-like design; the plant/object-like design got rejected due to the unnatural communication channel. Participants also reported that a nuanced solution would be preferred, i.e. both the look and the behaviour of the robot should not try to represent a person in all its characteristics, as the older adults would feel uncomfortable interacting with an artificial entity that tries to fully replace a human being.

The script for the first two focus groups covered questions about potential interests and concerns older adults may have about ECAs, provided brief definitions about robots and avatars, and gave examples of existing robots (NAO, iCAT, Flobi, FurHat, and XIBOT) as well as quality factors, desired functionality, appearance and aesthetics (design aspects), personality, motion, and movement features. The outcomes of the first two focus groups indicated that most participants preferred a robot presenting a friendly personality, but they did not feel the need for a robot having a backstory or a strong personality/character. The participants of the first two focus groups also emphasized the importance of maintaining social contacts with *real* humans, which is not replaceable by technology in their opinion. Most participants explained that they would not want an ECA to take over their tasks and responsibilities as this would deteriorate their abilities, unless they are no longer capable of taking care of such themselves. Instead, participants agreed that they would want their ECA to stimulate them to be independent for as long as they can (e.g. to be more active physically or to exercise their memory skills and cognitive abilities).

The third and fourth focus group were conducted with participants who were not highly educated, and who had severe somatic impairments. The participants in these focus groups were largely unfamiliar with robots and what a robotic technology is able to do. In fact, these participants were also unfamiliar with wireless and/or networked technology. As a result, it was difficult to discuss the possibilities of future technology with them – most of the time was spent on explaining to the participants what robot and networked technology is, and what it might offer for them in the future.

Based on the combined outcomes of all four focus groups, we conclude that participants think the usefulness of the robot, as well as its appearance, is strongly related to its purpose, features and functionality. Findings of the focus groups were consistent with related work by, e.g., Mitzner et al. (2010) and Wu et al. (2014): participants were mostly positive about what technology could do for them in terms of functionality, especially when their own capabilities start to decline [28,33]. The functionality of an ECA, in their opinion, varies based on individual preferences. In other words, it is essential to design for personalization and customization of the ECA. Able-bodied users (e.g. without any important impairments) emphasized that their requirements for interaction are very different from users with dementia or any communication impairments, such as aphasia. Still, common impairments in older adults, including hearing and visual impairments should be generally considered.

2) Functionality

The information obtained from the focus groups sessions aided the research team to select a final design: the astronaut-like with a spherical head and rectangular body. As input features, the robot would react to the users' movements and speech, detected through presence and proximity sensors and microphone. Additional functionality of the robot, in terms of output responses includes conversational features and companionship.

3) Implementation of the Bi-Bodied Conversational Agent

a) Robot body

To implement the robot body, the feedback of the end users (gathered through user studies) and the project constraints were considered (technical aspects, budget, time, etc.). Having a display for a face was discussed. This would have the advantage of it being easy to add extended screen-based functionality and interaction. However, finally, we opted for LED matrices as those would ensure discernibility and visibility of the displayed emotions, even in brightly lit rooms. White opaque material was used to visually connect the LED matrices. We combined these with moveable eyebrows to further enhance the perception of the digital agent as being physically embodied (as opposed to embedding the on-screen avatar in the robot head).

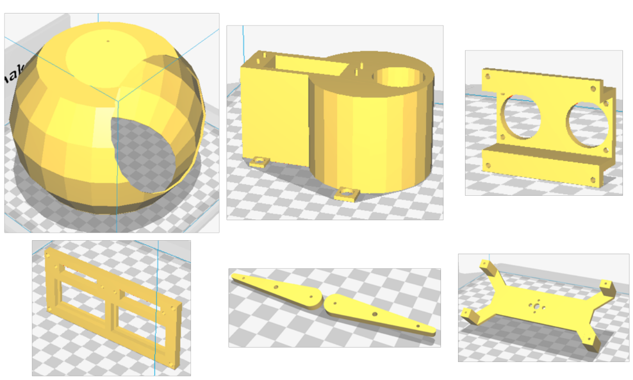


Fig. 3. Three-dimensional models of the head of the robot skeleton (skull, neck, proximity sensors, eyes, eyebrows and support piece.

Three-dimensional models of the robot skeleton were generated using SketchUp, integrating structural pieces (brackets for the eyes, eyebrow, neck, and skull) with external parts (robot body, eyebrows), and electronic components of the robot (Figure 3). A laser cutter was used to create the body box, with dimensions sufficiently large to house the electronics. The head skeleton, eyebrows, brackets and robot neck were 3D-printed in an UltiMaker 2. The components were assembled with screws to ensure a sturdy integration. Table 1 describes the list of components used in each robot part. Four Arduino Uno boards were used to implement and integrate the behaviours of the LED matrices (eyes and mouth), sensors, and neck movements.

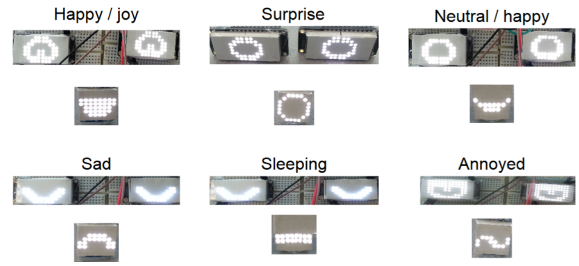


Fig. 4. Six facial expressions of the robot.

Based on the research conducted by Bennet & Šabanović (2014) [24], we decided to express the robot's emotions through facial expressions combining different movements for the eyebrows, eyes, and mouth (Figure 4). Also, bi-dimensional head movements were defined – horizontally (left and right movements), and vertically (up and down movements).

TABLE I
FINAL PROTOTYPE COMPONENTS OF THE ROBOT

Part	Material	Electronic Pieces
Head	- 3D printed globe - Paper mache finishing	
Eyes and Eyebrows	- 3D printed brackets	2 mini servo motors 2 8x16 LED matrixes (white) 1 Arduino Uno board
Mouth	- 3D printed brackets	1 8x8 LED matrix (white) 1 Arduino Uno board
Neck	- 3D printed structure - 1 tilt/pan bracket	2 normal-sized servo motors 1 Arduino Uno board
Body	- Laser cut rectangular box of white 3mm polyoxymethylene	1 LED display 2 proximity sensors 1 Arduino Uno board

b) Avatar body

A virtual avatar was created using the Crazytalk software by Reallusion [34] to mimic the behaviour and functionality of the physical robot. The image of the physical body of the robot was rendered to produce the virtual avatar using coordinate dissection, i.e. mapping anchor points of the image to a skeleton used for the automatic generation of facial expressions. Pictures of the avatar are presented in Figure 5.

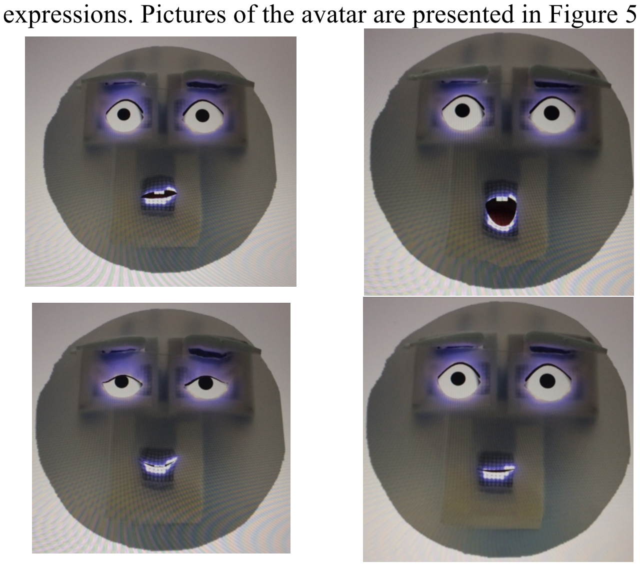


Fig. 5. The avatar, based on the robot, created with CrazyTalk

4) WoOz panel

We also developed a simple GUI for the Wizard-of-Oz to control either of the bodies, depending on which body is connected to the control panel (Figure 6). The whole panel was designed using MyRobotLab software [35]. Part of this control panel is also the functionality of "talking" to the user through a speaker set (independent of robot body), by typing text into the control panel and using text-to-speech synthesis software. ALICE 2.0 bot AIML scripts were used for partially automated conversation so as to relieve the Wizard from replying to all conversational utterances. The avatar and/or robot actions can be controlled by the Wizard-of-Oz using the WoOz panel as described in the study by Cheong (2011) [36].

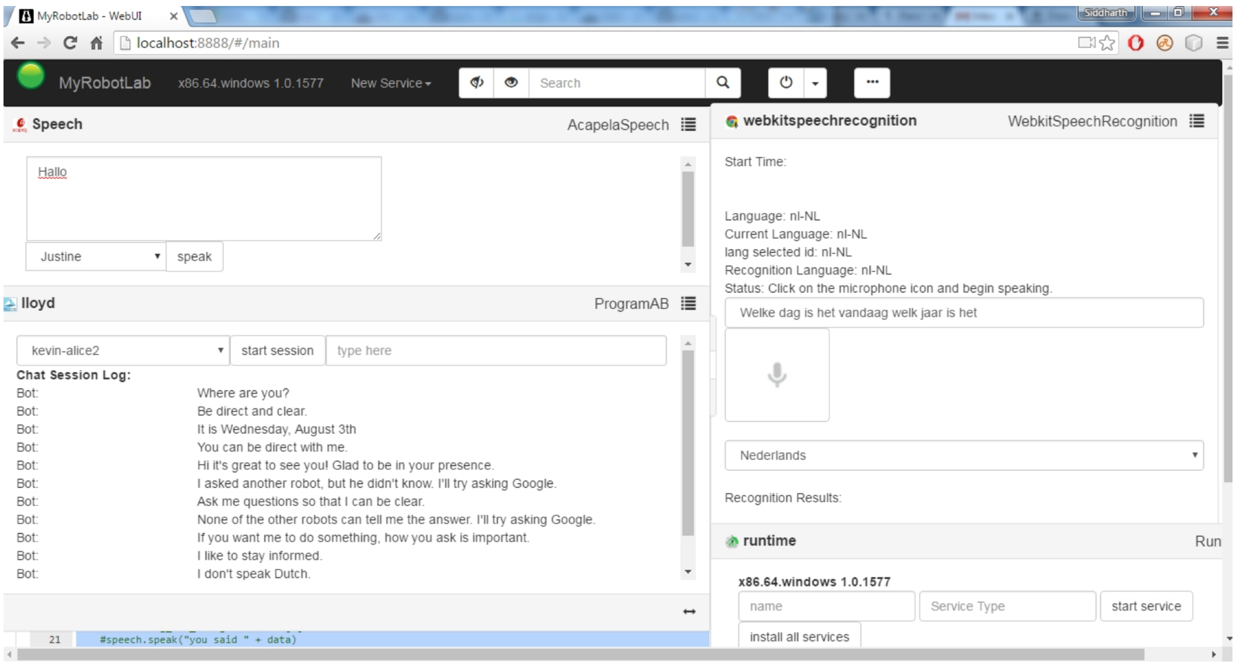


Fig. 6. MyRobotLab was used to develop the Wizard-of-Oz Panel

VI. DISCUSSION AND CONCLUSION

The key contributions of this work include: a conceptual framework for designing, evaluating and comparing two ECA formats (robot and virtual avatar); a computational component, integrating existing software and customized algorithms for speech recognition and conversational features; and a physical prototype, i.e. a robotic prototype for social communication activities with older adults. This study also provides insights in the perspectives of older adults upon embodied conversation agents (ECAs) as technological solutions for social communication. The current paper covers both design and implementation challenges to gain a deeper understanding of the needs and preferences of older adults when interacting with ECAs. Following the inclusive design principles we incorporated the potential users in all stages of design, development, and evaluation. Our results indicate that in principle older adults hesitate to adopt ECAs as technological solutions, for multiple reasons, ranging from: fear to lose or deteriorate their existing abilities and social contacts, lack of purpose or need, reliability issues and cost. Our results also show that older adults are more likely to accept the ECA technology when a clear need is identified and trust is built. In this study, we explored the feasibility of building and employing ECAs for the specified user population, setting the design foundations for future projects in the domain, and building a low-cost portable ECA solution to enable further evaluations (as described next).

As future work we plan to complement the evaluation of the technology created, exploring additional features of the robot, and gathering more users' insights for future design directions. We plan to refine the prototype built, optimizing and automating functions that are currently semi-automated. We plan to build a comprehensive documentation of the solution, to enable replication of the project by other research teams. Concerning the virtual agent, we plan to synchronize the updated version of the design with the current avatar, and as such conduct future evaluations and comparisons investigating the benefits, drawbacks, and specific advantages of a virtual agent vs. a physical agent, as perceived by older adults.

Creating and designing an avatar/robot for elderly users is a complex process that requires a lot of thought and experimentation. The combined approach of participatory design involving focus groups in different steps of the design process and micro-analysis of the users' interaction with the system has shown that the users – senior citizens – are not generally afraid of autonomous systems. In fact, we conclude that users' attitudes actually improve after interacting with the system.

The main lessons learned through this study is (a) how important it is to have a clear idea of the study population, as older adults may large vary on their needs and interests to have and/or use an ECA, and (b) that the research approach (sCE method [30]) used throughout this study is well-suited for guiding the design process, incorporating user requirements, and evaluating those requirements with the target group. This rapid development of incrementally more complex prototyping and evaluation may benefit other research projects, especially when designing for understudied groups of people, involving new and innovative technology, situated in specific and complex task domains and environments.

As design considerations for a system that would be socially acceptable, participants mainly stressed that the robot should be functional. Robot functionality is highly dependent on the target individual. This entails that a high degree of customizability is needed. Every individual needs a different design solution depending on their needs and possible impairments. Generally, people preferred a somewhat humanoid way of expressing emotions via a face, with individual people's preferences ranging from abstract to very human-like. More research is needed to determine the acceptability of the current system, as well as research determining if participants can recognize the emotions as programmed.

ACKNOWLEDGMENT

We thank all the participants who voluntarily took part in the study. We also thank all members of the Design Lab, University of Twente, Netherlands for help in designing the robot. Furthermore, we thank 4TU.Humans&Technology for funding the study.

REFERENCES

[1] Y. Barnard, M. D. Bradley, F. Hodgson, & A. D. Lloyd. (2013). Learning to use new technologies by older adults: Perceived difficulties, experimentation behaviour and usability. *Computers in Human Behavior*. *29(4)*, pp. 1715–1724. Available: http://doi.org/10.1016/j.chb.2013.02.006

[2] K. Zickuhr, & M. Madden. (2012). *Older adults and Internet use. Pew Internet*. Available: http://www.pewinternet.org/reports/2012/older-adults-and-internet-use.aspx

[3] A. D. Fisk, W. A. Rogers, N. Charness, S. J. Czaja, & J. Sharit. (2009). *Designing for older adults: principles and creative human factors approaches*. CRC Press.

[4] W. A. Rogers & T. L. Mitzner. (2016) Envisioning the future for older adults: Autonomy, health, well-being, and social connectedness with technology support. *Futures*. In press. Available: http://dx.doi.org/10.1016/j.futures.2016.07.002

[5] N. Selwyn, S. Gorard, J. Furlong, & L. Madden. (2003). Older adults' use of information and communications technology in everyday life. *Ageing and Society*. *23(5)*, pp. 561–582. Available: http://doi.org/10.1017/S0144686X03001302

[6] M. Ziefle, P. Pappachan, E. M. Jakobs, & H. Wallentowitz. "Visual and auditory interfaces of advanced driver assistant systems for older drivers". In: *Computers for Handicapped Persons (ICCHP)*, 2008. pp. 62–69. Springer. Available: http://dx.doi.org/10.1007/978-3-540-70540-6_8

[7] G. Lepicard, & N. Vigouroux. "Touch screen user interfaces for older subjects". In: *Computers Helping People with Special Needs (CHPSN)*, 2010. pp. 592–599. Springer. Available: http://dx.doi.org/10.1007/978-3-642-14100-3_88

[8] J. M. Beer, C. A. Smarr, T. L. Chen, A. Prakash, T. L. Mitzner, C. C. Kemp, & W. A. Rogers. "The domesticated robot: design guidelines for assisting older adults to age in place". In: *Human-Robot Interaction (HRI)*, 2012. pp. 335–342. ACM/IEEE. Available: http://dl.acm.org/citation.cfm?id=2157806

[9] S. Hanke, E. Sandner, A. Stainer-Hochgatterer, C. Tsiourti, & A. Braun. "The technical specification and architecture of a virtual support partner". In: *European Conference on Ambient Intelligence (AmI)*, 2015. CEUR. Available: http://ceur-ws.org/Vol-1528/paper4.pdf

[10] M. Kramer, R. Yaghoubzadeh, S. Kopp, & K. Pitsch. A conversational virtual human as autonomous assistant for elderly and cognitively impaired users? Social acceptability and design considerations. In: *Informatik*, 2013. pp. 1105–1119. Available: https://pub.uni-bielefeld.de/publication/2605000

[11] L. P. Vardoulakis, L. Ring, B. Barry, C. L. Sidner, & T. Bickmore. "Designing relational agents as long term social companions for older adults". In: *Intelligent Virtual Agents (IVA)*, 2012. pp. 289–302. Springer. Available: http://dx.doi.org/10.1007/978-3-642-33197-8_30

[12] R. Yaghoubzadeh, M. Kramer, K. Pitsch, & S. Kopp. "Virtual agents as daily assistants for elderly or cognitively impaired people". In: *Intelligent Virtual Agents (IVA)*, 2013. pp. 79–91. Springer. Available: http://link.springer.com/chapter/10.1007/978-3- 642-40415-3_7

[13] S. Thielke, M. Harniss, H. Thompson, S. Patel, G. Demiris, & K. Johnson. (2011). Maslow's Hierarchy of Human Needs and the Adoption of Health-Related Technologies for Older Adults. *Ageing International*. *37(4)*, pp. 470–488. Available: http://doi.org/10.1007/s12126-011-9121-4

[14] S. Kiesler, A. Powers, S. R. Fussell, & C. Torrey. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*. *26(2)*, pp.169–181. Guilford Press. Available: http://dx.doi.org/10.1521/soco.2008.26.2.169

[15] C. D. Kidd & C. Breazeal. "Effect of a robot on user perceptions". In: *Intelligent Robots and Systems (IROS)*, 2004, pp. 3559–3564. IEEE. Available: http://dx.doi.org/10.1109/IROS.2004.1389967

[16] K. M. Lee, Y. Jung, J. Kim, & S. R. Kim. (2006). Are physically embodied social agents better than disembodied social agents? The effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. *International Journal of Human-Computer Studies*. *64(10)*, pp. 962–973. Elsevier. Available: http://dx.doi.org/10.1016/j.ijhcs.2006.05.002

[17] J. Wainer, D. J. Feil-Seifer, D. A. Shell, M. J. Mataric. "Embodiment and human-robot interaction: a task-based perspective". In: *Robot and Human Interactive Communication (ROMAN)*, 2007. pp. 872–877. IEEE. Available: http://dx.doi.org/10.1109/ROMAN.2007.4415207

[18] J. A. Mann, B. A. MacDonald, I. Kuo, X. Li, E. Broadbent. (2015). People respond better to robots than computer tablets delivering healthcare instructions. *Computers in Human Behavior*. *43*, pp. 112-117 Elsevier. Available: http://dx.doi.org/10.1016/j.chb.2014.10.029

[19] L. Sucar, G. Azcarate, R. Leder, D. Reinkensmeyer, J. Hernandez, I. Sanchez, and P. Saucedo. (2009). Gesture therapy: A vision-based system for arm rehabilitation after stroke. In *Biomedical Engineering Systems and Technologies (BIOSTEC)*, *25*. pp. 531–540. Springer. Available: http://dx.doi.org/10.1007/978-3-540-92219-3_40

[20] E. Sandry. (2015). Re-evaluating the form and communication of social robots. *International Journal of Social Robotics*. *7(3)*, pp. 335-346. Springer. Available: http://dx.doi.org/10.1007/s12369-014-0278-3

[21] G. Mone. (2016). The edge of the uncanny. *Communications of the ACM*. *59(9)*, pp. 17–19. Available: http://dx.doi.org/10.1145/2967977

[22] M. P. Aylett, P. O. Kristensson, S. Whittaker, & Y. Vazquez-Alvarez. None of a CHInd: relationship counselling for HCI and speech technology. In *CHI*, 2014. pp. 749–760. ACM. Available: http://dx.doi.org/10.1145/2559206.2578868

[23] R. Aylett, M. Kriegel, I. Wallace, E. Márquez Segura, J. Mecurio, S. Nylander, & P. Vargas. "Do I remember you? Memory and identity in multiple embodiments". In: *Robot and Human Interactive Communication (ROMAN)*, 2013. pp. 143–148. IEEE. Available: http://dx.doi.org/10.1109/ROMAN.2013.6628435

[24] C. C. Bennett & S. Šabanović. (2014). Deriving Minimal Features for Human-Like Facial Expressions in Robotic Faces. *International Journal of Social Robotics*. *6(3)*, pp. 367–381. Springer. Available: http://dx.doi.org/10.1007/s12369-014-0237-z

[25] B. Görer, A. A. Salah, & H. L. Akın. (2016). An autonomous robotic exercise tutor for elderly people. *Autonomous Robots*. pp. 1–22. Springer. Available: http://dx.doi.org/10.1007/s10514-016-9598-5

[26] M. Volonte, S. V. Babu, H. Chaturvedi, N. Newsome, E. Ebrahimi, T. Roy, S. B. Daily, & T. Fasolino. (2016). Effects of Virtual Human Appearance Fidelity on Emotion Contagion in Affective Inter-Personal Simulations. *Transactions on Visualization and Computer Graphics*. *22(4)*, pp. 1326–1335. IEEE. Available: http://dx.doi.org/10.1109/TVCG.2016.2518158

[27] R. A. Paauwe, J. F. Hoorn, E. A. Konijn, & D. V. Keyson. (2015). Designing robot embodiments for social interaction: affordances topple realism and aesthetics. *International Journal of Social Robotics*. *7(5)*,

pp. 697-708. Springer. Available: http://doi.org/10.1007/s12369-015-0301-3

[28] Y. H. Wu, V. Cristancho-Lacroix, C. Fassert, V. Faucounau, J. de Rotrou, A. Rigaud. (2014) The Attitudes and Perceptions of Older Adults With Mild Cognitive Impairment Toward an Assistive Robot. *Journal of Applied Gerontology. 35(1)*, pp. 1–15. SAGE. Available: http://doi.org/10.1177/0733464813515092

[29] Blanson Henkemans, O. A., Van der Pal, S., Werner, I., Neerincx, M. A., & Looije, R. (2017, March). Learning with Charlie: A robot buddy for children with diabetes. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 406-406). ACM. Available: http://doi.org/10.1145/3029798.3036656

[30] M. A. Neerincx. (2011). Situated cognitive engineering for crew support in space. *Personal and Ubiquitous Computing. 15(5)*, pp. 445–456. Springer. Available: http://doi.org/10.1007/s00779-010-0319-3

[31] M. M. M. Peeters, M. Harbers, & M. A. Neerincx. (2016). Designing a personal music assistant that enhances the social, cognitive, and affective experiences of people with dementia. *Computers in Human Behavior. 63*, pp. 727–737. Elsevier. Available: http://dx.doi.org/10.1016/j.chb.2016.06.003

[32] Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—why and how. *Knowledge-based systems*, *6*(4), 258-266. Available: https://doi.org/10.1016/0950-7051(93)90017-N

[33] T. L. Mitzner, J. B. Boron, C. B. Fausset, A. E. Adams, N. Charness, S. J. Czaja, K. Dijkstra, A. D. Fisk, W. A. Rogers, & J. Sharit. (2010). Older adults talk technology: Technology usage and attitudes. *Computers in human behavior. 26(6)*, pp. 1710-1721. Elsevier. Available: http://dx.doi.org/10.1016/j.chb.2010.06.020

[34] "CrazyTalk 8". Available: http://www.reallusion.com Retrieved 2016-07-20

[35] "Myrobotlab". Available: http://www.myrobotlab.org Retrieved 2016-07-20

[36] W. L. Cheong, Y. Jung, & Y.-L. Theng. "Avatar: A virtual face for the elderly." In: *Virtual Reality Continuum and Its Applications in Industry*, 2011. pp. 491–498. ACM.

[37] N. Martelaro, V. C. Nneji, W. Ju, & P. Hinds. "Tell Me More: Designing HRI to encourage more trust, disclosure, and companionship". In: *Human Robot Interaction (HRI)*, 2016. pp. 181–188. ACM/IEEE. Available: http://dl.acm.org/citation.cfm?id=2906863

Dr. Marieke M.M. Peeters works at TNO as a research scientist artificial intelligence. Her research position is carried out within the Perceptual and Cognitive Systems group in Soesterberg. She combines her knowledge of both human cognition and artificial intelligence in her research. Her research focuses on human-agent interaction and human-agent relationships. She has a special interest in the application domains of safety & defence, education & training, and healthcare.

Vivian Genaro Motti is an Assistant Professor at the Department of Information Sciences and Technology at George Mason University (GMU) - USA. Dr. Motti holds a Ph.D. degree in Human-Computer Interaction (HCI) from Catholic University of Louvain - Belgium since 2013. At GMU, she leads the Human-Centric Design Lab. Her main research interests are HCI, UbiComp and Healthcare Informatics.

Helena Frijns is a MSc student of the Media Technology programme at Leiden University. In 2016 she obtained her BSc in Mathematics and its applications at Utrecht University and her bachelor degree in Fine Arts at the Royal Academy of Art in The Hague. Her current research interests include Human Computer Interaction and information visualization. She aims to combine doing research with making art.

Siddharth Mehrotra received the B. Eng. degree in Computer Science from the Galgotias University India in 2015. He is currently pursuing his MSc degree in Computer Science from International Institute of Information Technology, Hyderabad, India. His research interest lies in Music Cognition, Music Neuroscience, and Human Computer Interaction. In 2013, he joined the Department of Computer Science, Indian Institute of Technology, Guwahati, India as a Research Intern. He also served for Microsoft as Microsoft Student Partner from 2012 to 2015 and Member for India App Review board in 2015.

Tuğçe Akkoç received her BA from Yeditepe University with a double major in Psychology and Guidance & Psychological Counseling. She is an MA student of Cognitive Science at Bogazici University. Within the program she is tracking Psychology and Computer Engineering with a special interext in multimodal interaction. Currently, she joined the VR1 BAU Virtual Reality Lab as a researcher to find out the effects of VR exposure on the end users.

Sena B. Yengeç is an Electrical-Electronics Engineer. She has conducted research on an Environment Aware Social Agent in 2015. Her recent research involved work on Embodied Conversational Interfaces for Elderly Users.

Oğuz Çalık is an Electrical-Electronics Engineering, and International Relations student. His recent work involved work on designing and assembling a robot for elderly users.

Prof. dr. Mark A. Neerincx is a full professor in the II group and a principle scientist at TNO (the Netherlands Organization for Applied Scientific Research). His research interests include human-computer interaction in healthcare and space technology. His primary achievements are: (1) models for cognitive task load and emotion to support task performance and healthy lifestyle behaviour, (2) models of man-machine collaboration to adapt the type and level of support, (3) methods for cognitive engineering, and (4) a broad collection of best practices for usable and user-friendly cognitive systems.

First Time Encounters with *Roberta*: a Humanoid Assistant for Conversational Autobiography Creation

Minha Lee*, Stephan Schlögl†, Seth Montenegro‡, Asier López§, Ahmed Ratni**, Trung Ngo Trong¶, Javier Mikel Olaso§, Fasih Haider‖, Gérard Chollet††, Kristiina Jokinen¶, Dijana Petrovska Delacrétaz**, Hugues Sansen‡‡, María Inés Torres§

*Eindhoven University of Technology, the Netherlands
†Management Center Innsbruck, Austria
‡Hyx Ltd, United Kingdom
§Universidad del País Vasco (UPV/EHU), Spain
¶University of Helsinki, Finland
**Télécom SudParis, CNRS,Université Paris-Saclay, France
‖Trinity College Dublin, the University of Dublin, Ireland
††Intelligent Voice, United Kingdom
‡‡Shankaa, France

***Abstract*—During eNTERFACE we developed a dialog system design and conversation material for *Roberta*, an anthropomorphic assistant robot. The focus was on the first stage of what we call LifeLine dialogs, i.e. the conversational creation of users' life stories. Our goal is to help senior citizens record semi-autobiographical narratives while combating the deterioration of memory and speech abilities. We successfully completed modelling dialog scenarios for first time users. This allows *Roberta* to personalize future conversations based on each user's place of origin, work and education history, and hobbies, which are all information gathered during a user's first conversation with *Roberta*. We accomplished this through (1) an adaptable dialog system with topic management and multi-modal functionalities (specifically face recognition), by extending a RavenClaw-type dialog management framework, (2) using the Wizard of Oz (WOZ) data collection technique for categorizing introductory conversation material and identifying semantic rules and concepts, and (3) simulating the customization of a user's relationship with *Roberta* by introducing wizard-controlled contextual data, conversation history, and emotion awareness.**

Keywords*—dialog system, speech input, agent-based dialog management, topic trees, face recognition;

I. Introduction

The World Health Organization (WHO) reports that the number of citizens over the age of sixty is expected to double by 2050 [1]. People are living longer worldwide, but policies on healthy and meaningful ageing are yet to be fully realized, according to the WHO's comprehensive 2015 report on ageing and health [14]. Framing the elderly population as valuable contributors, rather than as stereotypically dependent individuals, will be beneficial to the entire society. Senior citizens' contributions to their communities are far more significant than care-related services that policy-makers invest in them [14]. Conversational storytelling, for example, promotes healthy memory and speech abilities, companionship, and a treasury of verbatim history for posterity. In line with this view, the *Roberta* project aims at constructing a robot listener which attempts to capture personal narratives through a LifeLine module; i.e. a feature which lets users converse about their personal lives.

The multidisciplinary field of human-robot interaction (HRI) is filled with ongoing developments. Robots are being heralded as new assistants of the ageing population by providing possibilities for companionship, health care assistance, and help for those who have difficulty with independent mobility. Yet, many practical challenges remain to be solved before robot companions become a reality.

At eNTERFACE 2016 hosted by University of Twente in the Netherlands, our focus was on feasibility. Many factors need to be incorporated for a functional HRI solution to best aid users (e.g. technical reliability, dialog design, context awareness, user experience, trust, etc.). We attempted to address a few of these challenges (i.e. dialog design, context awareness, as well as trust by researching how to best approach people; i.e. dialog initiation) for the HRI with *Roberta*. We worked on creating engaging conversational interactions and techniques for presenting non-human agents as conversation partners. *Roberta* should learn to adapt to users as a personal assistant with flexible speech capacity. *Roberta*'s LifeLine feature, which places great importance on natural dialog development, and was thus our signpost for those four weeks at eNTERFACE.

[1] http://www.who.int/mediacentre/news/releases/2015/older-persons-day/en/

While the project's long-term plan is for both users and *Roberta* to initiate topics for conversations, fight memory lapse and speech deterioration of users, and ultimately for *Roberta* to become a companion, not solely an assistant, our initial focus and task for eNTERFACE was on creating a dialog management system for intuitive interaction possibilities with *Roberta* for first time users. During the first meeting, *Roberta* should be able to set up a conversation history per user, ask for critical information about the user's life to begin a user's LifeLine in a conversational manner, and gather visual data from the environment, a face to be precise, to detect the user after his/her first encounter. In summary, the goal of this agenda was:

- to construct adaptable dialog system with topic management and multi-modal functionalities (specifically face recognition), by extending a RavenClaw-type dialog management framework;
- to use the Wizard of Oz (WOZ) data collection technique for categorizing introductory conversation material and identifying semantic rules and concepts; and
- so simulate the customization of a user's relationship with *Roberta* by introducing wizard-controlled contextual data, conversation history, and emotion awareness.

II. Background

To start, a Spoken Dialog System (SDS) is an assembly of natural language processing units that permits spoken interaction between a non-human agent or a knowledge base and a user. A SDS extended with a multimodal dialog system has tactile, biosensor, or visual data as input and output modalities. This requires further processing units to deal with sensor fusion.

Three tasks for dialog management are (1) contextually relevant interpretation of a user's speech (i.e. Natural Language Understanding), (2) the system's decision making process on an action to take (i.e. dialog Management), and (3) providing a suitable statement as a response (i.e. Natural Language Generation). The dialog manager is able to oversee the context of a discussion through the integration of identified user speech blocks alongside a system's own uttered messages in the conversation history.

A distinct user model is oftentimes created to keep track of user specific characteristics and his/her preferences. Additionally, a matching world model is provided to account for the general world knowledge and the surroundings. Conversation partners share information in order to establish a 'common ground' or discuss what is mutually known - a process which is known as grounding (i.e. to have a shared understanding of of the content which is discussed) and this constitutes an essential task of the dialog manager [2] [23]. Grounding requires conversation partners to be active in participating in a dialog so that they understand and recognize each other's contributions. Types of feedback may be verbal, gestural, or facial expressions.

The dialog manager decides on which action the system should perform at a given state in the dialog. Common dialog techniques are described in [9] where script-based, frame-based, agent-based, and statistical dialog management models are distinguished.

Script-based dialog management designs the dialog as a state diagram where each state and dialog action are paired. Although easy to design and implement for simple dialog tasks, they require a major effort to be built and maintained for complex dialogs. A frame-based DM provides more flexible dialog management by separating knowledge from dialog actions, so that the actions can be executed in various orders depending on the current information state (e.g. [24], [21]). The frame also provides a dialog context. However, more extended control algorithms and natural language understanding capabilities are necessary.

Many advanced dialog management architectures use an agent-based approach like the one employed by CMU's RavenClaw [1]. Originally developed in the Galaxy Communicator project, and further developed by [12], RavenClaw provides an advanced management structure with distributed software agents; i.e. dialog flow can be modelled by the software agents' reasoning about their own state and the next action. A recent work that demonstrates a sophisticated use of Ravenclaw is a framework called ReinForest, which extends Ravenclaw with a Semi Markov Decision Process in the dialog design structure for hierarchical reinforcement learning [28]. The dialog structure and reasoning can be flexibly designed, and it is also possible to take multimodal information and dynamic context into account. Similar architectures have been proposed for example in TRIPS [5], EDECAN [13] and JASPIS [25].

Interactive systems as a field is constantly developing, for applications that are fit for real-world user experiences are particularly challenging [4], [7], [20], and systems that are adaptive and multi-modal are especially demanding to build. Our attempt with *Roberta* was to construct a RavenClaw-based dialog system in order to support a systematic, step-by-step exploration and development process which should be continuously driven by new insights coming from different research efforts. The following sections further expound the respective efforts undertaken at eNTERFACE. First, Section III describes our efforts in collecting initial interaction and corpus data through a number of Wizard of Oz experiments. Next, Section IV discusses the system architecture *Roberta* is build on and Section V details the different dialog examples which we were able to construct based on the collected data. Finally, Section VI presents some concluding remarks and points to planned future research efforts.

III. Audiovisual dialogs through Wizard of Oz

Prototyping prior to building any system that is proficient in complex interactions requires tools and methods that consider future developments. An approach we used to test and conceive of *Roberta*'s capabilities is a low-fidelity solution called Wizard of Oz (WOZ) [3]. The WOZ technique allows researchers and designers to gather user feedback and interaction on a system prototype that typically requires significant effort to implement. In a 'WOZ study' a human researcher, the 'wizard', controls a system's envisioned functionalities before the entire system is built. Spoken interactions as high-end prototypes are costly to fully develop in terms of time and resources. Thus, WOZ studies are cost and time effective measures to obtain interaction data and data on possible dialog pathways taken by users.

A successful WOZ experiment needs (1) a prototyping framework for a realistic study environment, in that the wizard has to be hidden from participants so that they believe that they are interacting with the actual system and (2) pre-made scenarios that showcase features of the envisaged system solution to be tested. For *Roberta* we were able to fulfill both requirements during eNTERFACE 2016. The significant point in this study was to test how participants responded to *Roberta* on different categories of conversations, such as on work and hobbies, through WOZ. We cover this aspect in more detail in section 3 and 5.

A. The WOZ Prototyping Tool

The WebWOZ prototyping platform we used can be found on Github[2], and is specifically aimed at designing interactions based on natural language processing [18]. Its web-based state means that existing language technologies like automatic speech recognition and text-to-speech synthesis units can be easily integrated. We collected interaction data with 15 participants through a pre-configured version of WebWOZ that runs on a VirtualBox system image [19]. The system relies on Google Speech API[3] for speech recognition and on Open-Mary TTS engine[4] for speech synthesis.

Our data collection used the below described scenarios as a guideline but allowed the wizard to be flexible when needed. Thus, we prepared possible utterances that the wizard could make the system say to a participant. For unexpected participant responses, the wizard provided utterances ad hoc, and these ad hoc instances were considered and incorporated as needed in building the dialog manager. Such realistic interactions gave us new considerations for strategizing potential future dialogs, specifically on how designing conversational agents must better prepare for unplanned responses or questions from participants; people may be expecting more from a non-human entity, be it humorous responses or insightful comments, as conversations get longer. Even though the greater goal for *Roberta* is to offer open domain exchanges, the first stage of development was directed towards limited scenarios that mimics initial conversations between two strangers.

B. WOZ Driven Roberta Scenarios

An interdisciplinary team is paramount to creating user scenarios that cater to users' actual needs rather than stereotyping potential users, according to [6]. Our team consisted of individuals with diverse skills and roles, from engineers to human-computer interaction researchers. Consequently, also our ideas as to how the interactions with *Roberta* should be handled, were deviating, which eventually made us design intake conversations that are broad, yet unambiguous enough for a natural first time encounter. *Roberta*'s topic categories were (1) introduction, (2) professional life, (3) hobbies, and (4) conversation closing. These are departure points for construing a cohesive autobiography about a user through conversations, and this is the main goal of the LifeLine module that *Roberta* will be eventually able to offer.

As mentioned by Dahlbäck et al., real-life conversations are essential for building natural language dialog systems since HRIs are filled with unanticipated conversational directions [3]. We captured users' sociodemographic information in the intake conversation because this is critical data that assistive robots need to be aware of in order to cater to their conversation partners [6]. Yet, many researchers, designers, and engineers do not place importance on gathering and using sociodemographic data for user personalization [6]. Our WOZ experiments demonstrated that it is possible for *Roberta* to gain relevant sociodemographic information per user during the first conversation.

Roberta starts every encounter by seeing if a user's face is already in the system's database, which is the beginning of the introduction topic category. If a face is not recognized, *Roberta* first asks for permission to add a new user to her database. Secondly, *Roberta* inquires the user for his/her name and the spelling if it is a name that is not yet easily recognizable. Thirdly, *Roberta* asks about a user's place of origin and whether or not it is the same as where he/she now lives. Lastly, *Roberta* asks for age, although our studies showed that this can be a sensitive question. Therefore, if an answer is not offered, *Roberta* does not re-ask the question, and saves the question for a future encounter. *Roberta* can contribute to this introduction scheme by providing answers about herself, such as "I am Roberta", "I am infinite", and "I'm from Enschede". These possible answers make data gathering conversational. Below are two sample conversations from our study that show the introduction step.

Conversation 1

Roberta: Hi I am *Roberta*. Do you want to register and talk with me?
User: Yes I would like that.
R: What is your name?
U: Jorge.[5]
R: Where are you from?
U: Nicaragua. And you?
R: I'm from Enschede. Do you know it?
U: Yes I do, but I still prefer Nicaragua.

Conversation 2

R: Hi I am *Roberta*. Do you want to register and talk with me?
U: Yes, I am here to talk to you. I am glad to know you.
R: Good. What is your name?
U: I'm Sungahn from Korea.
R: Can you spell that for me?
U: S-u-n-g-a-h-n.
R: Great. Where do you actually live?
U: Sorry, I don't understand your question.
R: Where do you live?
U: Ah, you mean where do I now live. I live in Rotterdam.
R: Do you like it there?

[2]https://github.com/stephanschloegl/WebWOZ
[3]https://cloud.google.com/speech/
[4]http://mary.dfki.de/

[5]All names and places have been changed to protect participants' identities.

U: It's okay. It's nice, but can be challenging.

C. Moving to LifeLine Dialogs

The pre-determined topics of discussions were based on *Roberta*'s LifeLine module, which aims to converse with a user about his/her life [17]. This module opens up an endless number of possible topics to converse about within a rather simple paradigm of questions and answers. We hope that the elderly population will find *Roberta* as a willing listener of past stories as *Roberta* gets to know more about each user. This helps to strengthen a user's memory recall as well. LifeLine at a meta-level can link various people's lives together as their stories converge in *Roberta*'s system, cross-referencing stories and enriching details of relevant events, which can function as a more private social network for individuals who share specific events that *Roberta* incorporates. Users' shared stories can be departure points of future discussions with *Roberta*. Yet, to this end it is important to note that data security and the preservation of privacy is an important and valuable (future) research topic – one which is, however, out of scope for the current stage of our efforts.

With a minimal amount of information gathered during the initial meeting between *Roberta* and a user, we saw an emergence of a rich picture of a person. The intake conversation is critical, for it starts a user's LifeLine for a semi-autobiographical approach to spoken narrative built around a dialog with *Roberta* acting as a *ghostwriter*. *Roberta* organizes a user's life stories around a unique time line as she builds on following conversations with each user. Of equal importance is categorizing a user's data with the first conversation since this aids in creating tailored dialog approaches per user. Even though our participants at eNTERFACE were mainly academic researchers and university students, our WOZ experiment gave us building blocks for future research direction.

There were a total of 15 participants who talked with *Roberta*. We asked for basic information on users' name, age, place of origin, past educational history, work experience, and hobbies. This gives a rich picture of a user even with an initial meeting. What we identify as the biggest challenge is in accommodating participants' responses that are possibly meant to be humorous or playful. Two excerpts are shared below.

Conversation 3

R: What do you do?
U: I bother people usually.
R: Why did you choose to do that?
U: I did not choose it. This is my nature.

Conversation 4

R: Do you have any hobbies?
U: It sounds like you're speed dating. I like surfing, like playing a guitar.
R: So I guess you know a lot about that.
U: Speed dating or...?

Most participants took *Roberta*'s questions or statements at face value, which makes standard responses prepared by researchers work as intended. Yet, when participants introduce unexpected utterances as in Conversations 3 and 4 above, the naturalistic flow of a conversation may be interrupted. Conversation 4 demonstrates the difficulty of going beyond an "interview" or "speed date" scenario since the intention of the first conversation is to mostly gain information from participants in order to later personalize conversations to fit each user. This is not a challenge that is only unique to *Roberta*, for many non-human agents designed to be interactive partners cannot account for all conversational nuances users introduce. However, we acknowledge that more research is necessary in order to adjust conversations to account for these ad hoc utterances. We plan to conduct further WOZ experiments with a more general population, especially the target group of elderly users.

IV. System Architecture

The system we built includes a perception input layer, an understanding module, and an interaction module which encompasses the dialog manager. The dialog manager controls the user blackboard that handles the dialog history and an open domain interactive agent. The agent takes information from a digital repository of 'system knowledge' (based on past conversations) and forwards it to the output layer responsible for the generation of the corresponding text utterances. The goal is to integrate a humanoid robot to the aforementioned components in the next phase of the project.

A. Perception and Understanding

a) Automatic Speech Recognition: – Using a local and remote modules, our Automatic Speech Recognition (ASR) was processed in a distributed fashion. We used Intelligent Voice[6] and Google[7] for remote access to provide rich linguistic data and knowledge processing. A local module consisting of an HMM/DNN decoder that runs on NVIDIA GPGPU should help as a backup. Our language of choice for this first stage was English.

b) Audiovisual Signal Processing: The SudFrog (Telecom SudParis Face Recognition open-source reference software[8] was incorporated for audio-visual signal processing. It is part of the open-source reference systems first implemented by the BioSecure Network of excellence [15]. The system has been widely tested on various large scale and challenging biometric databases, for example, the FRGC [16] and the MOBIO [11] databases. Recently it has been implemented in an audio-visual person verification application running on an iPad [26].

At eNTERFACE 2016 we used SudFrog to add an additional input channel, which should help us contextualize interactions. That is, a number of our study participants enrolled into a dialog with *Roberta* with their faces. The enrollment phase consisted of face and landmark detection, geometric normalization, and the creation of a template with Gabor features [22]. This template represented the subject for further comparisons. In case of a second encounter with the same participant,

[6]http://www.intelligentvoice.com/
[7]https://cloud.google.com/speech/
[8]http://share.int-evry.fr/svnview-eph/

Roberta would check whether she would 'recognize' this face, and if yes would provide the dialog manager with the name of the participant.

c) Natural Language Understanding: – Semantic parsers based on statistical classifiers outclass semantic decoders that manually construct semantic grammars for Natural Language Understanding (NLU), as shown in [8]. Full capabilities of such a semantic parser requires training on a large number of sentences. Thus, a main goal of our eNTERFACE efforts was to collect a respective dataset of sentences, and transform them in corresponding grammar rules and concepts using the Phoenix Semantic Parser [27]. This semantic parser is not a statistical classifier; it is rule based. The designs need to be hand-crafted.

We were able to collect data from interactions with 15 participants, which gave us some initial constructs and vocabulary to work with. Functions used in the parser are "name", "place", "yes", "no", "age", "ask feeling", "ask age", "ask name", "ask place", "and you", "ask occupation", "occupation", "occupation academic", "no occupation", "work topic", "work place", "work passion love", and "work passion hate", as shown in Figure 1. The grammar rules need to be broadened with the introduction of vocabulary from more subjects and situations, and this can be achieved by collecting more data with varied settings in terms of concepts and interactive elements of *Roberta*.

B. Interaction and Dialog Management

Our system uses RavenClaw for independent agent-based dialog management, a framework that is task-independent and plan-based. Through a dialog agents tree, tasks are indicated at each level for domain specific logic [1]. The dialog specification task plans out interactions as a hierarchical structure made up of sub-dialogs. A set of rudimentary dialog agents are found at terminal points of the tree. The main idea behind RavenClaw is simple "slot-filling". User data is modelled in advance, and the system requests additional data in a predefined order during the conversation. A user provides this data via spoken input, and the semantic parser serializes the input stream using semantic mapping (i.e. it uses key-value pairs) to fill in each data slot. In order to complete this task effectively, RavenClaw introduces a multi-thread based agents system which can robustly handle a great number of use-cases and which is tolerant to faults or errors occurring during the conversation. All the logic that lies behind the DMS is coded into these agents. There are four classes of agents:

- Inform Agent: prompts input from the system to the user using Natural Language Generation (NLG) and Text To Speech (TTS).
- Request Agent: expects a specific type of information from the user and is activated after the semantic parser receives a given semantic label.
- Execute Agent: calls back-end operations, which allows for an open dialog system to handle tasks that are higher in complexity.
- Expect Agent: is a listener that runs in the background, only active with specific user input and capable of processing unforeseen user input.

Based on these agents, the non-terminal points in the tree constitute sub-tasks or sub-dialogs. Agencies depend on preconditions, triggers, and success and/or failure criteria. The whole dialog can be considered as finite-state machine, which traverses from a beginning state among different branches to reach an end state.

The limitation of this approach is that the script files become larger as the logic grows, which makes it difficult to maintain or add new features. Figure 1 illustrates the overall process of building the DMS. Themes per user such as hobbies or family life can be added to our DMS, as we have covered with WOZ interviews; categories in Figure 1 are not exhaustive, but serve as starting points.

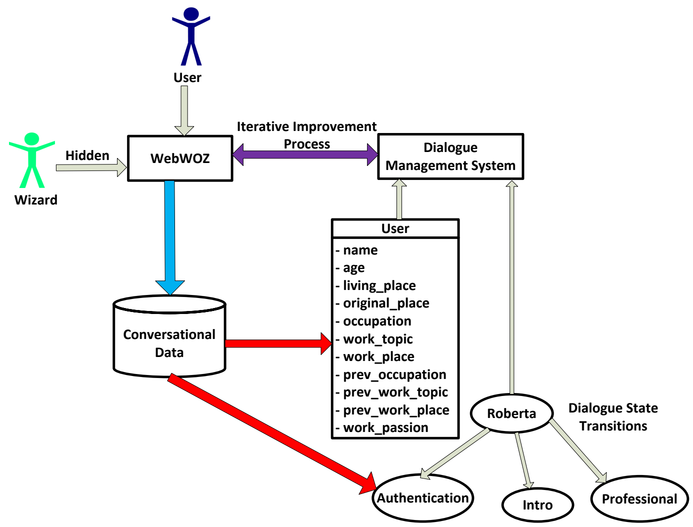


Fig. 1. The process of constructing the DMS.

In sum, Ravenclaw is adaptable for countless disciplines and applications [1]. Its strength is in concepts that dictate information transfers down the tree, described in detail in [1]. We hence used RavenClaw for its flexibility in its topic dependent dialog manager and furthered the framework's capacity; the integration with the WebWOZ Wizard of Oz prototyping platform was handled via the Olympus Hub system.

C. Output

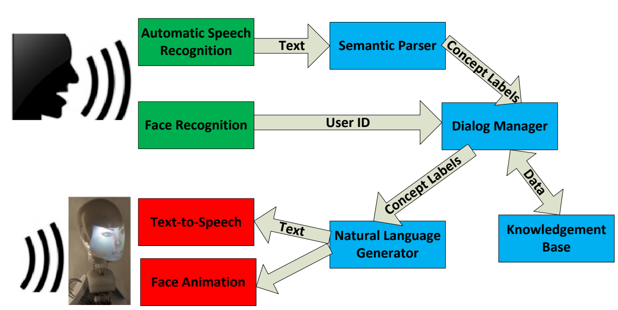


Fig. 2. Main modules of *Roberta* developed at eNTERFACE.

The main components of the *Roberta* dialog framework developed at eNTERFACE are depicted in Figure 2. The ASR produces a text transcription of the input speech from the user, that goes through the semantic parser into concept labels. These labels and the user ID produced by the Face Recognition module feed into the Dialog Manager, that outputs other concept labels, according to input labels and user ID, and this also supplies the user data in the Knowledgement Base. The output concept labels were used by the Natural Language Generator for the output sentence of *Roberta*. This was used to generate *Roberta*'s speech via a TTS, and also for facial animation.

We utilized Unity's game engine for synchronous facial animation and text-to-speech output. We used a 3D model called Liza by GR Modular System, which comes with with meshed viseme animations[9]. Thus, we were able to take time-saving measures because this method includes pre-loaded animations for facial expressions and pre-generated viseme articulations. In doing so, the phonemes provided by MaryTTS text-to-speech instances were mapped to matching visemes for English and in sync with 3D animations. *Roberta* was able to speak fluently by instantiating the same text-to-speech engine as the phoneme server to the dialog manager; the integration of Unity 3D engine and ActiveMQ as the shared messaging provider pushed messages read by the Linux build of *Roberta*'s Unity project.

V. Dialog Examples

Figure 3 displays the topic-based dialog structure from our WOZ conversations. An agent was defined for each topic based on the WOZ scenarios. The interaction dialogs gathered served as a guide for designing the agents. Our flexible dialog structure relies on concepts that act as triggers that control agents or subtree execution, which results in topic change as often as needed. For example, if a user brings up a hobby, this triggers *Roberta*'s interest in this hobby. She will search in the knowledge repository to check for any relevant information about this hobby, or if the hobby is recognized in the user blackboard, or the dialog blackboard history as an attribute value, then *Roberta* is able to ask the user about his/her hobby via a request agent.

The WOZ corpus became the basis to develop the "authentication" and the "intro" agencies, which can be seen in Figure 3. Additionally, *Roberta* can inquire about a user's hobbies or professional activities as shown in the tree in Figure 3. If a user first discusses his/her professional life, *Roberta* can ask about is/her current job, job tasks, an average workday, and his/her education. If a user does not have job, *Roberta* then requests information on what a user used to study or do.

Conversation 5
R: What's your job?
U: I'm now working on an airport security project.
R: What tasks do you do?
U: That's very specific. Right now I am working on voice activation.
R: Tell me about your typical workday.
U: Well I wake up around eight and come to work.
R: Wonderful. What did you study?
U: I studied electrical engineering in Germany.

Roberta then ventures to talk about hobbies. If a user does not have any hobbies, *Roberta* will ask about music, sports or traveling as specific categories. All three categories have anticipatory questions and/or statements. For instance, *Roberta* can ask "do you like music?", and then "who is your favorite musician?" or "what is your favorite genre?" or "do you play any instruments?".

Conversation 6
R: What is your hobby?
U: I don't have a lot of time for hobbies.
R: Do you like music?
U: Yes I like music.
R: What is your favorite genre?
U: I like classical music.
R: Do you play any instruments?
U: I used to play the guitar when I was very young, but not anymore.

Conversation 7
R: What is your hobby?
U: I like photography and traveling.
R: Tell me more about the places you've visited.
U: I visited so many countries... Japan, Thailand, South Africa, Costa Rica...

The end of the conversation can also be controlled by trigger conditions, which are shown in Figure 3. As a transition, *Roberta* is able to share a joke and/or get permission to meet again. *Roberta* then bids farewell to her conversation partner.

R: Do you want to listen to a joke?
U: Yes please.
R: Why was 7 afraid of 9?
U: Why?
R: Because 789.
U: Very funny.
R: Do you want to talk more some other time?
U: Yeah maybe some other time.
R: Talk to you later.
U: Bye.

Our WOZ experiment participants expressed a tendency to ask questions to *Roberta*. To accommodate this, we added an "and-you" agency, which used some expectation agents to make the dialog more natural, and also information agents that give *Roberta* possible answers (cf. Figure 4). Simply put, "and-you" agency means *Roberta* can ask, as well as answer, conversational "small talk", as aforementioned in Section 3-B, WOZ Driven Roberta Scenario. For instance, she can say that she is well and that she is from Enschede, the Netherlands, in case participants ask for such information.

Lastly, we supplemented *Roberta* with a capability to give feedback as a preliminary design, which allows her to respond to a user. Our WOZ study showed that how *Roberta* responds to participants is vital, motivating our need for generating feedback. An example of standard feedback is in Conversation 4, a statement such as "I guess you know a lot about that", which is a general remark. We found that participants did

[9]https://www.assetstore.unity3d.com/en/#!/content/52234

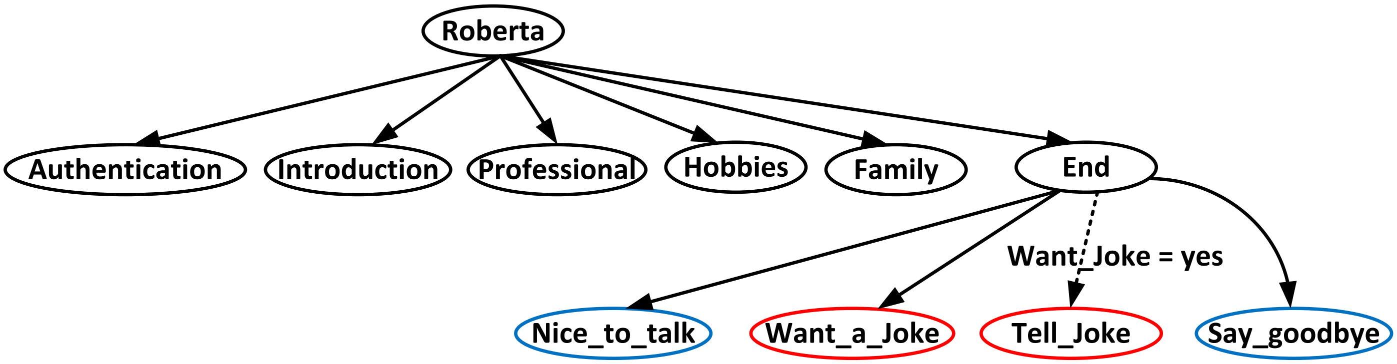


Fig. 3. Topic-based dialog structure.

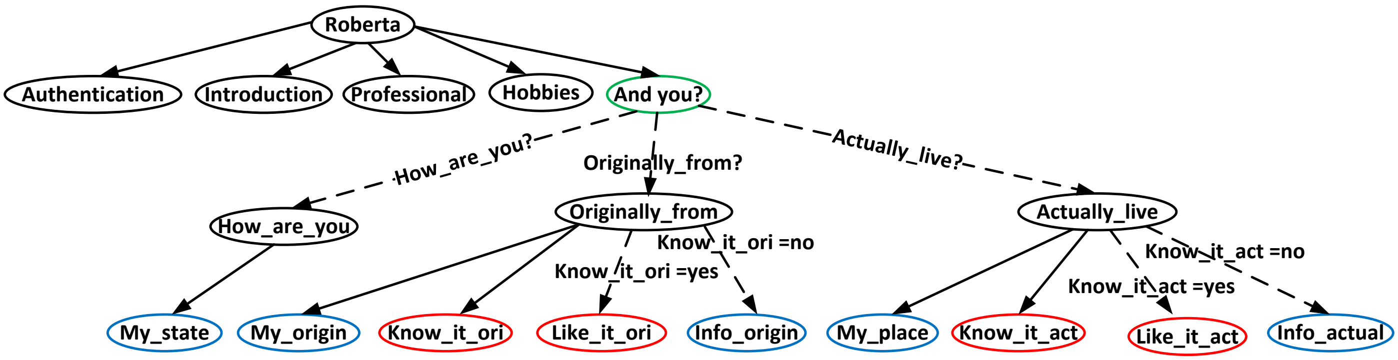


Fig. 4. The dialog structure for when users ask questions.

provide informative answers on name, age, place of origin, and work related topics, but as conversations get lengthier, participants tend to deviate from our expected dialog flow. This means that by the time participants reached further categories we wanted to expand on, such as hobbies, the expectation is a more equal contribution from the agent, beyond the ability to ask questions and give standard responses. Thus, feedback statements that are achieved with inform and execution agents, which make a sufficiently flexible dialog possible, need further expansion based on further research.

VI. Concluding Remarks and Future Work

We have described research undertaken thus far on an agent-driven dialog data collection for a dialog system design that is fit for *Roberta*, a personal assistant humanoid robot. At eNTERFACE 2016, our efforts were predominantly on developing *Roberta*'s speech capabilities and the dialog design of LifeLine, a module for conversational creation of a user's autobiography with a robot. We achieved this with data from WOZ sessions that consist of 15 videos of conversations and the corresponding audio data. This corpus was our foundation for developing functional speech understanding, face recognition, dialog task and natural language generation modules. The entire structure currently includes a spoken dialog system that is flexible and extends the pre-existing RavenClaw dialog management framework with multi-modal capacities and topic management.

However, *Roberta*'s LifeLine module at this point serves as a proof-of-concept and will be further developed. We particularly require more interaction data to expand the context *Roberta* should be able to deal with, and corpus data to improve speech recognition and language understanding. Also, we need to move the system out of a protected lab setting into people's natural interaction environments. Thus, similar WOZ experiments will be pursued to compose an audiovisual database that includes studies conducted in more realistic surroundings, such as in houses of elderly people, and we will further include other languages through tele-assistance support. This should not only help collect the necessary corpus data to tune language technologies but also expand our understanding of the different discussion topics *Roberta* needs to be prepared for. In conclusion, real-life introduction of assistive technologies that are intended to be "human-like" in communication capabilities to build rapport with users must design for greater flexibility in conceptualizing dialog scenarios, even for a specific use-case such as conversational autobiography creation. The LifeLine Dialogs project proposed for eNTERFACE 2016, still ongoing, was presented in FELT2016 workshop and will be published in the conference proceedings [10].

Acknowledgments: The authors thank the organizers of eNTERFACE'16 and the University of Twente. We also acknowledge the Spanish Ministry of Science under grant TIN2014-54288-C4 and "ADAPT 13/RC/2106" project (http://www.adaptcentre.ie/) at the School of Computer Science and Statistics, Trinity College Dublin, the University of Dublin, Ireland for making this work possible.

References

[1] Dan Bohus and Alexander I Rudnicky. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332–361, 2009.

[2] Herbert H. Clark and Edward F. Schaefer. Contributing to Discourse. *Cognitive Science*, 13(2):259–294, 1989.

[3] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pages 193–200. ACM, 1993.

[4] Maxine Eskenazi, Alan W. Black, Antoine Raux, and Brian Langner. Let's Go Lab : a platform for evaluation of spoken dialog systems with real world users. In *InterSpeech*, 2008.

[5] George Ferguson and James F. Allen. Trips: An integrated intelligent problem-solving assistant. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, pages 567–572, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.

[6] Priska Flandorfer. Population ageing and socially assistive robots for elderly persons: the importance of sociodemographic factors for user acceptance. *International Journal of Population Research*, 2012.

[7] Fabrizio Ghigi, Maxine Eskenazi, María Inés Torres, and Sungjin Lee. Incremental dialog Processing in a Task-Oriented Dialog. In *InterSpeech*, pages 308–312, 2014.

[8] James Henderson, Paola Merlo, Ivan Titov, and Gabriele Musillo. Multilingual Joint Parsing of Syntactic and Semantic Dependencies with a Latent Variable Model. *Computational Linguistics*, 39(4):949–998, 2013.

[9] Kristiina Jokinen and Michael McTear. *Spoken dialog Systems*, volume 2. Morgan & Claypool Publishers, 2009.

[10] Asier López, Ahmed Ratni, Trung Ngo Trong, Javier Mikel Olaso, Seth Montenegro, Minha Lee, Fasih Haider, Stephan Schlögl, Gérard Chollet, Kristiina Jokinen, Dijana Petrovska-Delacrétaz, Hugues Sansen, and María Inés Torres. Lifeline dialogues with roberta. In Jose F Quesada, Francisco J. Martin-Mateos, and Teresa Lopez-Soto, editors, *Future and Emerging Trends in Language Technologies, Machine Learning and Big Data, FETLT2016*, Lecture Notes in Artificial Intelligence, page to appear. Springer International Publishing, 2017.

[11] Christopher McCool, Sebastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matejka, Jan Cernockỳ, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635–640. IEEE, 2012.

[12] Javier Mikel Olaso, Pierre Milhorat, Julia Himmelsbach, Jerome Boudy, Gérard Chollet, Stephan Schlöogl, and María Inés Torres. A Multi-lingual Evaluation of the vAssist Spoken dialog System. Comparing Disco and RavenClaw. In *International Workshop on Spoken dialog Systems*, 2016.

[13] Javier Mikel Olaso and María Inés Torres. dialog system based on EDECÁN architecture. In *Lecture Notes in Artificial Intelligence*, pages 547–551. Springer, 2010.

[14] World Health Organization. *World Report on Ageing and Health*. World Health Organization, 2015.

[15] Dijana Petrovska-Delacrétaz, Gérard Chollet, and Bernadette Dorizzi. *Guide to biometric reference systems and performance evaluation*. Springer, 2009.

[16] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005.

[17] H. Sansen, M. I. Torres, G. Chollet, C. Glackin, D. Petrovska-Delacrétaz, J. Boudy, A. Badii, and S. Schlögl. The roberta ironside project: A dialog capable humanoid personal assistant in a wheelchair for dependent persons. In *2016 2nd Int. Conf. on Advanced Technologies for Signal and Image Proc. (ATSIP)*, pages 381–386, 2016.

[18] S. Schlögl, G. Doherty, and S. Luz. Wizard of Oz experimentation for language technology applications: Challenges and tools. *Interacting with Computers*, 27(6):592–615, 2015.

[19] S. Schlögl, P. Milhorat, G. Chollet, and J. Boudy. Designing language technology applications: A wizard of Oz driven prototyping framework. In *Proceedings of the EACL Conference of the European Chapter of the Association for Computer Linguistics*, pages 85–88, 2014.

[20] Manex Serrras, Naiara Pérez, María Inés Torres, and Arantza Del Pozo. Entropy-Driven dialog for Topic Classification: Detecting and Tackling Uncertainty. In *International Workshop on Spoken dialog Systems*, 2016.

[21] Mark Ter Maat and Dirk Heylen. *Flipper: An Information State Component for Spoken dialog Systems*, pages 470–472. Springer Berlin Heidelberg, 2011.

[22] Massimo Tistarelli, Manuele Bicego, José L. Alba-Castro, Daniel Gonzàlez-Jiménez, Mohamed-Anouar Mellakh, Albert Ali Salah, Dijana Petrovska-Delacrétaz, and Bernadette Dorizzi. *2D Face Recognition*, pages 213–262. Springer London, London, 2009.

[23] David R. Traum. A computational theory of grounding in natural language conversation. Technical report, University of Rochester, Rochester, NY, USA, 1994.

[24] David R. Traum and Staffan Larsson. *The Information State Approach to dialog Management*, pages 325–353. Springer Netherlands, Dordrecht, 2003.

[25] Markku Turunen and Jaakko Hakulinen. Jaspis-a framework for multilingual adaptive speech applications. In

InterSpeech, pages 719–722, 2000.

[26] Alexander Usoltsev, Dijana Petrovska-Delacrétaz, and Houssemeddine Khemiri. Full video processing for mobile audio-visual identity verification. In *Int. Conf. on Pattern Recognition Applications and Methods ICPRAM 2016*, 2016.

[27] Wayne Ward et al. The CMU air travel information service: Understanding spontaneous speech. In *Proceedings of the DARPA Speech and Natural Language Workshop*, volume 1, pages 127–129, 1990.

[28] Tiancheng Zhao. ReinForest: Multi-Domain Dialogue Management Using Hierarchical Policies and Knowledge Ontology. Technical report, School of Computer Science, Carnegie Mellon University, 04 2016.

Social Communicative Events in Human Computer Interactions

Kevin El Haddad[1], Hüseyin Çakmak[1], Marwan Doumit[2], Gueorgui Pironkov[1], and Uğur Ayvaz[3]

[1]TCTS lab - University of Mons - Belgium
[2]iCES Corporation - Virginia - USA
[3]Information Systems Engineering - Mugla Sitki Kocman University - Turkey

***Abstract*—In this paper we report on our work towards contributing to the Social Signal Processing and to the Human-Computer Interaction fields. It is a research focused project aiming at processing so called social Communicative Events (SCE) which we define as a subset of social signals. We present initial work on automatic SCE detection and classification, laughter processing and neutral to smiled speech conversion. To do this we first review some of the related freely available databases. We then show promising results but with still a lot of room for improvement.**

***Index Terms*—Social Signal Processing, Social Communicative Events, Affective Computing, Laugh, Smile, Synthesis, Recognition, Machine learning, Deep Learning.**

I. Introduction

THE social signal processing (SSP) field is dedicated to study multimodal expressions and signals we use during our social interactions [1], [2]. These signals can be understood and interpreted by others and linked to a certain semantic meaning. They can be expressed using facial expressions, speech, nonverbal expression, paralinguistic expression, postures, gaze, etc... SSP can be beneficial for applications of various fields, in particular human-computer interaction (HCI) applications. This project aims at contributing to the social signal processing field, in order to improve HCI systems. But in order to simplify the problem, we focus only on paralinguistic and nonverbal expressions expressed audibly or through facial expressions and we will refer to this subset of social signals as Social Communicative Events (SCEs). Nodding the head is an example of an SCE. It is a nonverbal expression communicating agreement to the interlocutor. Another example would be smiling while talking. Indeed, it can have emotional (expressing amusement) as well as social functionality (e.g. being polite) [3].

Among the most popular of social signal expressions in the current research studies are emotional/affective expressions. Emotion and sentiment processing in Human-Computer Interactions (HCI) systems is becoming popular nowadays. Indeed, taking these aspects of communication into account would make the interactions more human-like and therefore more comfortable for the users. Several work can be found concerning emotion, affect and sentiment processing [4]. The most common cues considered for emotional expression detection or recognition are facial expressions [5], [6] and the voice [7], [8], [9], [10]. Work can also be found targeting expressions rather than emotions. Indeed some expressions can be used to express different emotions or affects. This is the case for laughter and smiling for example. Previous research work focused on such expressions rather than emotions. In [9], [10], [11], mono and multimodal laughter detection systems are presented for example. Laughter synthesis was also the subject of previous work [12], [13], [14]. In general, emotions are expressed through multimodal or unimodal expressions. But, in a dialog, multimodal expressions can also be used to communicate ideas rather than emotions. In fact, work can be found on detection of head movements expressions (noding, and head shaking)[15], [16] and on backchannel and feedback classification [17], prediction [18] and generation [19]. As mentioned previously, this project's goal is to contribute to the SCE processing field, and more particularly to the synthesis and the recognition areas of this field. As will be seen in the following section, SCEs, other than laughter and smiles, are rather rare and often not well annotated in current the open source and accessible databases. So a first task for this project, with the goal of contributing to the SCE recognition domain, is to tackle the problem of automatic detection and classification of multimodal SCEs in database.

Also in the SCE recognition domain, since laughter detection systems can already be found and apparently give satisfying results, two task were defined to process detected laughs. The first one is to build an automatic laughter component classification system and the second one concerns a laughter intensity estimation system. In the SCE synthesis domain, given the data available (as will be explained in the corresponding sections) we present our initial work on neutral to smiled speech conversion.

Given the nature of the tasks defined, we propose machine learning based approaches. In fact, the SCE detection task can be seen as an unsupervised learning problem, the laughter processing tasks are classification tasks which are well handled by machine learning algorithm, and deep learning and statistical parametric modeling techniques proved to be very efficient for speech conversion tasks.

So, data is required. After preliminary investigations, obtaining SCE data turned out to be a challenging task. That is why the first week was dedicated to tackling the problem of collecting and analyzing data. Instead of recording the data, we rather rely on the existing databases since a considerable amount of them are available. So, in what follows, the data collection and analysis will be presented in Section II. In this sections, we also propose a short review of the databases considered. We also explain in more detail the choice of the tasks previously mentioned. Each of the four defined tasks are then presented in separate sections. We finally conclude and give perspectives on our future work in Section VII.

II. Data Collection and Analysis

An analysis of the available databases was necessary in order to define specific machine learning tasks related to SCE synthesis and recognition. In this section, we will first report of the selected databases content with respect to SCEs. We will also explain in more detail the tasks defined mentioned in the introduction.

A. Databases Summary

Since SCE can be expressed in different modalities (e.g. nodding is a visual expression, laughter and smiles are audio-visual expressions, backchannels are audio expressions, etc...), both the audio and visual modalities were considered during the data collection step. The databases considered for this project are listed in Table I. Considering the limited period of the workshop, it was important that the data we worked with had specific criteria:

- The databases should contain annotations that would help locating interesting SCEs.
- They should be open source.

Table I gives a summary of the databases considered with characteristics such as the modality of the data, the nature describing whether the data collected were acted or natural, the available annotated SCE and finally whether a segmentation of the speech is available or not.

Database	Modality	Nature	SCE	Speech
AMI* [20]	A/V	Nat./Act.	L	Yes
ICSI [21]	A	Nat.	L	Yes
le CID [22]	A/V	Nat.	L	Yes
Uva-Nemo [23]	V	Nat./Act.	S	No
AVLC [24]	A/V	Nat.	L	No
AVLASYN [25]	A/V/M	Nat.	L/S	No
SL-db [12] (French)	A	Act.	L/S	Yes
SM-db [26] (French)	A	Act.	S	Yes
IEMOCAP* [27]	A/V/M	Act.	L/E	Yes
SAVEE [28]	A/V/M	Act.	S/E	Yes
CCDb* [29]	A/V/M	Nat.	L/S/E	No

TABLE I
Databases considered during data collection with corresponding characteristics (SL-db correspond to Speech-Laugh database and SM-db to Speech-Smile database). Concerning the modality: A, V and M refer to Audio, Video and Motion Capture respectively. Concerning the SCEs: L, S and E correspond to laugh, smile and Emotions (e.g. happiness, anger, sadness, etc...) respectively. * indicates databases with annotations for body/head movement SCEs.

Concerning the AMI and ICSI, breath labels could be found in the annotations, but they were not considered as an SCEs in this project due to the ambiguous interpretations they could have, i.e. many of them most probably do not have any semantic or emotional meaning and thus they cannot be considered as SCE.

It seems like unlike other SCEs laughs and smiles are frequently taken into account in databases annotations. Other SCEs (head nodding, understanding/non-understanding signals, confusion expressions, backchannels etc...) although present, are neglected in these databases annotations. This might be due to there importance in social communications [3] and the relative homogeneity of these expressions among speakers compared to other SCEs (i.e. laughter or smiling are more like to have a common pattern than the disgust expression for example).

B. Tasks Definition

For the reasons mentioned in the introduction and considering the available data, the following tasks were defined to contribute to the SCE synthesis and recognition domains:

1) Automatic detection and classification of multimodal SCEs;
2) Laughter component classification;
3) Laughter intensity estimation
4) Neutral to smiled speech conversion

The first three tasks are related to the SCE recognition domain while the last one is related to the synthesis one. Indeed the previous data analysis indeed showed that SCEs other than laughs or smiles, although present in some databases, are often neglected and rarely well annotated. Thus task 1) was defined. It will not only help increase the available SCE data, but it will also be a first step towards SCE classification. Indeed further work would then use these systems to analyze the detected expressions in order to classify as precise types of SCE. For this a we present initial work on non-verbal expression detection systems in the audio and visual cues using Voice Activity Detection (VAD) and Facial tracking systems. Indeed because of preliminary investigations, we expected the previously mentioned databases to contain SCEs that could be used for future synthesis or recognition systems but which were not annotated.

Considering the frequent presence of laughter in currently available databases (such as in the AMI, CCDb, ICSI and le CID databases), and due to the fact that work can already be found on automatic laughter detection systems (see Introduction), we decided to define tasks 2) and 3) that would help taking advantage of these laughs for further processing. Indeed, concerning 2), the laughter pattern as defined in [30] is formed of sequences of unvoiced (usually fricatives) and voiced (usually vowel-like) sounds. The unvoiced and voiced parts can each have several classes and the point of this task is to be able to classify each incoming laughter frame the most accurately as possible. Such an automatic annotation system would be very useful for laughter synthesis [13] systems for instance since it would bypass the tedious task of manually annotation the laughter samples. In fact this tool will help us obtain a much larger annotated database that would be used to produce more naturally sounding synthesized laughs.

Concerning task 3), a laughter intensity/arousal estimation system could be used to improve the automatic annotation of the laughter data by tagging the laughs with an estimated arousal intensity level. This intensity information could be further used for example to control the level of intensity of a synthesized laugh or to extract more information from the detected laugh. Indeed, it was shown that laughter intensity can inform on several characteristics of the speaker such as his social status for instance [31].

Smiles are also available and annotated in the databases above in both the visual (facial expression recorded as video or motion capture) and audio (speaking with smiling) cues. A task defined here was more focused on the audio cue. Considering this cue, transcribed smiled speech can be found in the Speech-smile and the SAVEE databases. The data available are, for most of them, acted and recorded in "clean" conditions which make them well suited for analysis or synthesis tasks. In these two databases for each smiled sentence, a corresponding neutral sentence can be found from the speaker. That is why we defined task 4) concerning neutral to smiled speech conversion. Obtaining a system capable of converting any neutral speech to a smiled while sounding natural, will help improve HCI systems by adding more naturalness to the agent's speech.

Further work related to the visual cue of smile intend to improve the current state of the art of the smile detection systems (for both the audio and visual cues). But Considering the limited period of the workshop and limited amount of participants, this was not one this project's objectives.

III. Automatic SCE Detection in Databases

As mentioned in Section II-A, although present in the databases, some SCEs (such as audio, visual or audiovisual backchannels, fillers, etc...) are neglected in the annotations. And this might be due to the variability of certain emotional or social expression. This variability makes it harder to use these data to build a recognition, detection or synthesis system. Here, we want to tackle this problem by proposing and initiating work on detecting and classifying semantically the SCEs used in daily social communication - SCEs like the sounds and/or facial expressions we make when we understand or get confused during a conversation. In other words we want to be able to first detect and then "understand" or extract information from these SCEs. Here we propose a first step on an SCEs detection system that works in the audio and visual cues using a VAD and a facetracker with an automatic landmark detection. In the audio cue, if we know the position of the speech segments from the annotations, we can detect the other sounds and classify them a specific type of SCE or as "noise" or "trash" (this could be done using the speakers voices as references for example). A facetracker would help us detect a subject's face in the visual cue and then an automatic landmark detection would help us differentiate a neutral face from any other. The recognition can also then be made between 3 classes: smile, laughter and other expression since smile and laughter detection already exist. With these detection systems in both cues separately and/or combined, we would be able to extract SCE from data and analyze them for further applications. Ultimately, we expect to have a system that would classify SCEs with respect to the semantic information they can provide.

A. VAD Evaluation For SCE Detection

Here the silence removal method [32] was tested on a task of detection for voiced/unvoiced short audio events also considering their position with respect to speech. For this, 25 non-verbal audio events were considered from the CCDb. This database was chosen because it contains quasi-naturalistic dialogues and thus natural SCEs. Also it contains annotations for backchannels, surprise and confusion expressions. These events came from 3 different speakers (one session per speaker) and can be classified as voiced/unvoiced events and as either backchannels, fillers, laughs or affect bursts. These classes were chosen as preliminary assessments of this method's efficiency. The SCE were annotated manually and the VAD was applied on the whole session each time. For each SCE, the mean error between the predicted and the annotated beginning and ending times limiting the SCE was calculated per speaker in seconds and the results are given in Table II.

Speaker	Number of SCE	Mean Error (sec.)
Speaker 1	7	1.82
Speaker 2	10	0.92
Speaker 3	8	1.69

TABLE II
VAD times estimation errors with respect to annotated time limits of SCEs from the CCDb

Please note that some of the SCEs were located close to speech in the database and in that case, the VAD considered it as part of the sentence. This considerably increases the mean error values presented n the table above. This can be tackled in future work by simply using the speech segments annotations as reference and deducing the SCEs real limits by subtracting the times. Other VADs will also be tested in future work for this task.

B. Facetracker and Automatic Landmark Detection

One of the difficulties of dealing with SCEs, is that some expression can be vary greatly from one speaker to another on the audio cue, visual cue or even both. So, concerning the visual cue, our objective was to obtain a tool capable of automatically labeling, for a specific video frame, each face detected with an ID with respect to their position in the video. This would help us to correspond a certain audio event detected in the audio cue to a specific facial expression detected in the video cue. So, a facetracker with an automatic landmark detection system was implemented using the OpenCV library [33]. A first step into detecting non-neutral facial expression from subjects in data, is to be able to label the detected faces with IDs for an efficient automatic annotation. So, using the position of the detected faces in a specific video, we are able to attribute them an ID based on the order they appear on a horizontal axis: the face detected the further on the left of the axis will be labeled as 1 and the nth face to the right will be labeled as n.

The fusion of both the audio and visual cues should improve even more the results concerning the SCEs detected.

IV. Laughter Component Classification

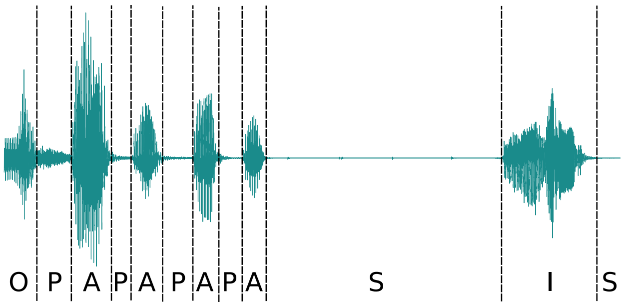


Fig. 1. Automatic laughter annotation. O = laughter Onset, P = Pulse, A = Vowels A, S = Silence, I = Inhalation sound

This work is part of a larger framework of automatic laughter annotation. We present here our work during this workshop on automatic classification of laughter frames. The goal here is to classify laughter frames as accurately as possible. Indeed, an accurate classification of these frames would help us obtain an accurate segmentation of a single laugh into several laughter components as shown in Fig. 1. As explained earlier, a laugh is formed of sequences of unvoiced (fricatives) and voiced (vowel-like) sounds, each of them can be broken into several classes. Here, the following classes were defined:

1) Nasal
2) Plosive
3) Grunt
4) Fricative exhalation
5) Nareal fricative inhalation
6) vowel o
7) Nareal fricative exhalation
8) vowel a
9) vowel
10) silence
11) Glottal Stop
12) Fricative Inhalation

A supervised learning approach was used for this task. First, annotated laughter was required. That's why the data was taken from the AVLASYN and AVLC databases (see Table II-A. The data is composed of laughter triggered by funny stimuli. Laughs were annotated for 1 male from AVLASYN and 1 male and 1 female from AVLC.

All the 334 laughs were divided into segments using a window width and non-overlapping shift of 10 ms. We obtained a total of 85000 frames. These were split into 75% training and 25% testing data.

Several deep learning architectures were considered for this task such as autoencoders, Long Short-Term Memory (LSTM) and Bidirectional LSTM [34], [35], [36].

Surprisingly, the best result is obtained using a simple Multilayer Perceptron of 3 hidden layers with 56, 128 and 56 nodes respectively. The training was made using the Adam algorithm [37]. Since our data are primarily formed of voiced/unvoiced and the voiced are divided in different vowels, the features chosen for training the system were simple 13 order MFCCs and pitch. These were extracted using a 30 ms wide window and a 10 ms window shift.

Fig. 2 give a heatmap of the accuracy of this MLP for this frame classification task for each of the 12 classes previously mentioned. Then integers on the horizontal and vertical axes represent the classes. The class labels are on the vertical axis and a darker color indicate the a higher percentage of correct classification. In overall this MLP gave us 75% of accuracy.

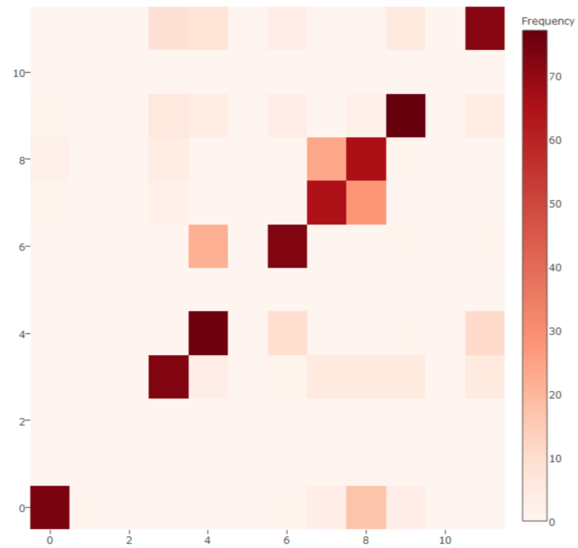


Fig. 2. Heatmap for MLP classification accuracy

We can see from this graph that most of the classes are very well classified, but others not. This may be caused by a poor representation of the data the features used offered. Please bare in mind that these results are obtained using simple features. Indeed we did not have time to investigate other features but this will be one of our goals in future work. The energy or dynamic features for instance, might help discriminate between different unvoiced data.

V. Laughter Arousal Estimation

In this section we present an ongoing work on a laughter intensity (arousal) estimation system which is also part of the recognition/detection tasks. This estimation system will be used in the future to extract more information from laughter and enrich the semantic content of a detected laugh. Intensity is an important dimension of laughter. The motivations behind this is related to the importance of the intensity (arousal) characteristic concerning laughter. The notion of intensity seems so natural that most researchers do not define it (e.g., [38], [39], [40]). In [41], Ruch defines the emotion of exhilaration, which is one of the emotions leading to laughter. He discusses different levels of intensity of this emotion and the corresponding behaviors, from smile at low intensity to laughter accompanied by posture changes (throwing back the head, vibrations of the trunk and shoulders) at high intensity. Furthermore, intensity is encoded differently by individuals, with reference to their own laughing style [40].

Since intensity is a fundamental dimension, frequently and naturally used to describe laughs, it appears as an important feature to be able to estimate for further use in laughter synthesis [42] or recognition. Indeed, it can give valuable information about the state of a participant in a human-machine interaction system. It is also a convenient layer in interactive systems to separate the processes of deciding to laugh (with a target intensity), which is independent from the laughter synthesis voice and style, and synthesizing the corresponding laugh, which obviously depends on the modeled individual traits. To avoid confusion with acoustic intensity (amplitude, loudness, etc.), in this report we will use the term *arousal* to refer to the intensity of the emotional state leading to laughter. Arousal is frequently used to characterize affects, for example in the 2-dimensional valence-arousal plane.

Here we give an overview of this ongoing work aiming at using laughter arousal efficiently as a parameter in human-agent interaction systems. First, an online arousal annotation experiment was carried to build a database. This database is then used to obtain first results towards laughter arousal estimation.

A. Online perception tests

To collect the arousal data, online arousal tests were conducted. Participants were asked to rate the arousal of laughs on a 5-point scale ranging from 0 to 4. Laughs from 3 subjects were evaluated in this test. Two subjects (1 male, 1 female) from the AVLC Database [24] and one subject (male) from the AVLASYN Database [25]. A total of 334 laughs were used. The number of laughs from each of the subjects is given in Table III.

AVLC DB (Subject 6)	67
AVLC DB (Subject 14)	65
AVLASYN DB (D4)	202
TOTAL	334

TABLE III
NUMBER OF LAUGHS FOR EACH SUBJECT IN THE EXPERIMENT

Due to relevant data availability, the laughs from the AVLC Database were evaluated only on the audio while laughs from the AVLASYN database were evaluated along 3 different modalities ; audio only, video only (video without sound) and both together.

In the case of the AVLASYN Database, 7331 ratings were collected from 226 participants (135 males and 91 females from 18 to 77 years old with an average age of 31.46 and a standard deviation of 10.23). The pipeline followed for the test is as follows :

1) First a registration form is filled by the participant. In this form, we ask different information about the participant : name, email, gender, if headphones or loudspeakers are used.
2) The next page of the test gives information about the task that will be asked and the context.
3) In the whole test, each participant is asked to give 36 evaluations divided in three parts. Each part is related to a modality (audio, visual or audiovisual). The order in which the 3 parts are presented is randomly defined when the user registers. In the first part, 12 files are successively shown to the participant. The files are randomly picked from the available 205 files with the constraint that a file that has already been shown to a given user can not be shown again. When the first part is finished, the other parts present the same laughs as those chosen randomly for the first part. Each time an evaluation is validated, it is saved so that if a participant decides to quit the test before finishing, the given evaluation are not lost.
4) When the 3 parts of the test are finished, a final form is proposed for participants who want to leave a comment.

Table IV gives the average number of time each file has been evaluated in each of the 3 parts.

Modality	Average (std)
Audio only	11.78 (3.47)
Video only	12.03 (3.30)
Audiovisual	11.95 (3.48)

TABLE IV
AVERAGE NUMBER OF TIME EACH FILE HAS BEEN EVALUATED IN EACH PART OF THE TEST

In the case of the AVLC Database, 1505 evaluations were collected from 40 participants (32 males and 8 females from 20 to 61 years old with an average age of 35.38 and a standard deviation of 10.43). The pipeline followed for the test is the same as above but it contains only one part which is the audio only and each participant was asked to evaluate 40 laughs. Each file has been evaluated 11.40 times on average with a standard deviation of 2.96.

B. Audio laughter arousal estimation

The previously described databases was then used in an arousal prediction/estimation task of a given audio laughter file.

To do this, we propose here to use a Gaussian Mixture Model (GMM) based approach. First, silences are removed from the input audio laughter files. Then, a set of features are extracted from these files and the features and the a subset of these features were created based on the correlation these features have with the arousal level. The ones that were the most correlated with the arousal level were kept. The selected features are then used to train GMMs with full covariance matrices. Doing so, we can model the relationship between the input acoustic features and the corresponding arousal levels. The GMM mapping framework used in this work was first introduced in 1996 by Stylianou [43] for voice conversion. The implementation used here is the one of Kain [44] also used in recent work such as [45].

1) Feature selection: The set of features extracted can be mainly divided into:

- Cepstum features which are widely used in emotion, speaker and speech recognition tasks.
- Prosodic features (such as the pitch)
- Temporal features (extracted from the temporal domain signal, like the energy)

Some features are scalar values related to the whole file in the first place while others are continuous features extracted using 10ms windows and 25ms frame shift. Some of the features were extracted using the Straight [46] vocoder. Due to the continuous nature of these features, statistical descriptors were extracted from them in order to represent the signal. As mentioned previously the correlation between these features and the corresponding arousal score obtained from the previously described online experiment was considered in order to select some of these features. Among the most correlated features, we mainly find F0 related features, Chroma vector related features, the mean of the zero-crossing rate and energy entropy standard deviation.

2) Results: The data were split between training, testing and validation data. The features subset obtained was used to train the GMMs which were then tested using a leave one speaker out protocol. The results obtained from these evaluation were promisingly showing accuracy scores as high as 85%.

VI. Neutral to Smiled Speech Conversion

Part of the synthesis generation tasks is a voice conversion application concerning the paralinguistic event of smiled speech. In this task, the goal is to be able to convert neutral speech to smiled speech while obtaining the most natural sounding result possible. This task was previously attempted in previous work [47], [26]. In this work we intend to ameliorate our previous results by improving the naturalness of the speaker using Deep learning methods. Our methodology for this is to use sentences read in two styles of speech: neutral and smiled. The neutral and its corresponding smiled sentence will be used to train a DNN to convert the former style to the latter. So we used the Speech-smile db (see Table I). This database contains sentences recorded from the same speaker reading them in both neutral and smiled styles. A more detailed description can be found in [26]. We thus have approximately 1 hour worth of each speech style. Three types of DNN architecture were considered for this tasks:

- MLP: Due to the seemingly simple task, a simple architecture is expected to give good results.
- Autoencoders: it gave good results with tasks involving denoising speech signals which is similar to our task
- LSTM: this architecture seem to have good results when it comes to model signals with sequential/temporal characteristics.

During the workshop, the conversion task was tested only on MFCCs. These were extracted using the SPTK toolkit [48] and a 10 ms sliding window of 30 ms width. Since the duration of neutral sentence and its corresponding smiled sentence are not the same, dynamic time wrapping (DTW) was used on the MFCC frames to temporally align every neutral-smile pair on the phoneme level.

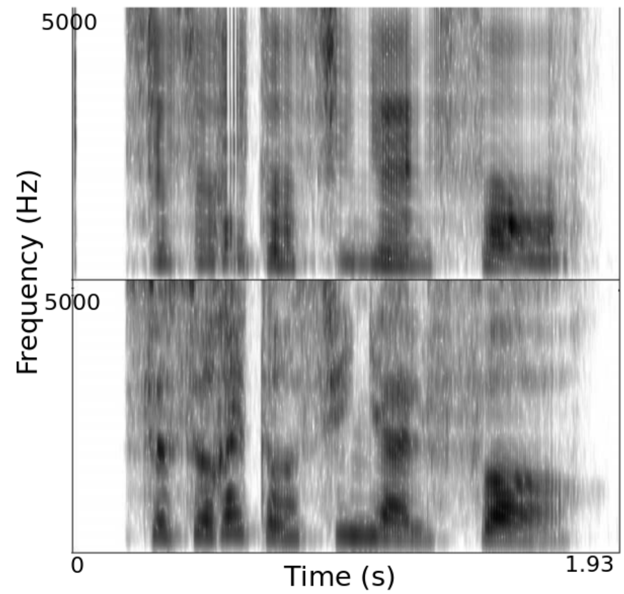


Fig. 3. Spectrogram comparison between input (bottom) and output (top) of an autoencoder.

Unfortunately no result good enough for synthesis were obtained till now using these setups but the most encouraging results were obtained using autoencoders. For this, several configurations were tried but in general most of them had four encoding layers and three decoding layers (if we count the bottleneck in the encoding layer part). For all the work described in this section (not just for the autoencoders) the data were split in a 80-10-10% for training, testing and validation. The best results till now were obtained by pre-training the autoencoders using the only the neutral speech in the database and the fine tuning it with the neutral and smiled speech. To explain this more, we first pre-train the autoencoder, using the neutral speech features frames as the input and the same frame as the target output. This was done using all the neutral training data for 8000 epochs in a mini-batch of 200 batches approach. The network was then fine-tuned using the neutral features as the input and the smiled speech features as the target. This was done with 5000 epochs and the same mini-batch approach.

For the sake of analysis Fig.3 shows the spectrograms after synthesis using the autoencoders after pre-training and before fine tuning. This was done using the MLSA filter of the SPTK toolkit. The one on the bottom correspond to the original spectrogram and the one on the bottom to the one obtained after prediction. As we can see these spectrograms are not exactly the same. It seems that the predicted one is less precise at the position of some formants. Indeed, as we can see, slices of the original spectrogram are more intense than others, while the corresponding slices for the predicted one seems more uniform. This might be due to the fact the amount of the input features is not enough for the neural network to learn a representation of the input data. In the near future we intend to find a better architecture configuration and use a better set of features. The STRAIGHT software [46] is also considered for this application in future work.

VII. Conclusion and Perspectives

In this report we presented a research-oriented project attempting to contribute to the social signal processing field and in particular the SCE synthesis and recognition areas of this field. After reviewing and investigating the available databases, we concluded that the laughs and smiles were frequently taken into account in the annotations while other types of SCEs were very often neglected. From this investigation, we proposed in this report four tasks we initiated during this workshop to tackle the problem of automatic SCE annotation, improve laughter processing by proposing an automatic laughter component classification system and a laughter arousal estimation system, and finally we proposed adding a paralinguistic aspect to neutral speech which is smiling.

The work presented here were all first approaches, all leaving room for a lot of improvement. Considering first the SCE automatic annotation task, unsupervised deep learning techniques would be a good way to build a fully functioning system. Indeed deep learning has proved its efficiency for similar tasks in the past. We plan on improving the automatic laughter annotation system by using more efficient features and adding other DNN architectures to our experiment. We also plan to re-implement or develop a laughter detection systems which will be coupled with this annotation system and the laughter arousal estimation system presented here. This will give us a fully function laughter processing pipeline. Finally concerning the neutral to smile conversion problem, we intend to use Stacked Denoising Autoencoders instead of simple autoencoders as was proposed in [49]. Recurrent Neural Networks might also be used for the time alignment problem between neutral and smiled sentences instead of the DTW used here.

Acknowledgment

We would first like to thank the eNTERFACE committee and the eNTERFACE'16 organizers for allowing us to organize this project. This work was partly supported by the Chist-Era project JOKER with contribution from the Belgian Fonds de la Recherche Scientifique (FNRS), contract no. R.50.01.14.F.

References

[1] A. Pentland, "Social signal processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, July 2007.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743 – 1759, 2009, visual and multimodal analysis of human spontaneous behaviour:. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608002485

[3] K. El Haddad, H. Cakmak, S. Dupont, and T. Dutoit, "Laughter and Smile Processing for Human-Computer Interactions," in *Just talking - casual talk among humans and machines*, Portoroz, Slovenia, 23-28 May 2016.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98 – 125, 2017.

[5] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ser. ICMI '04. New York, NY, USA: ACM, 2004, pp. 205–211. [Online]. Available: http://doi.acm.org/10.1145/1027933.1027968

[6] J. Kumari, R. Rajesh, and K. Pooja, "Facial expression recognition: A survey," *Procedia Computer Science*, vol. 58, pp. 486 – 491, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915021225

[7] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320310004619

[8] M. R. Amer, B. Siddiquie, C. Richey, and A. Divakaran, "Emotion detection in speech using deep networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3724–3728.

[9] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144 – 158, 2007.

[10] T. Neuberger and A. Beke, *Automatic Laughter Detection inSpontaneous Speech Using GMM–SVM Method*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 113–120.

[11] S. Scherer, F. Schwenker, N. Campbell, and G. Palm, "Multimodal laughter detection in natural discourses," in *Human Centered Robot Systems*. Springer, 2009, pp. 111–120.

[12] K. El Haddad, S. Dupont, J. Urbain, and T. Dutoit, "Speech-laughs: An HMM-based Approach for Amused Speech Synthesis," in *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, Brisbane, Australia, 19-24 April 2015, pp. 4939–4943.

[13] H. Çakmak, "Audiovisual laughter synthesis - a statistical parametric approach," Ph.D. dissertation, University of Mons, February 2016.

[14] J. Urbain, "Acoustic laughter processing," Ph.D. dissertation, University of Mons, 2014.

[15] A. Kapoor and R. W. Picard, "A real-time head nod and shake detector," in *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, ser. PUI '01. New York, NY, USA: ACM, 2001, pp. 1–5. [Online]. Available: http://doi.acm.org/10.1145/971478.971509

[16] Y. Chen, Y. Yu, and J. M. Odobez, "Head nod detection from a full 3d model," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015, pp. 528–536.

[17] C. Navarretta and P. Paggio, "Classification of feedback expressions in multimodal data," in *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, 2010, pp. 318–324.

[18] L.-P. Morency, I. de Kok, and J. Gratch, *Predicting Listener Backchannels: A Probabilistic Multimodal Approach*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[19] S. Fujie, K. Fukushima, and T. Kobayashi, "Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system." in *INTERSPEECH*, 2005, pp. 889–892.

[20] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, ser. MLMI'05. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 28–39.

[21] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," 2003, pp. 364–367.

[22] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle," *Traitement Automatique des Langues*, vol. 49, no. 3, pp. 1–30, 2008. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00349893

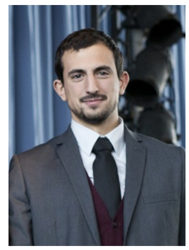
[23] H. Dibeklioğlu, A. A. Salah, and T. Gevers, *Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 525–538.

[24] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The avlaughtercycle database," in *Proc. of the Seventh Int. Conf. on Language Resources and Evaluation (LREC'10)*, 2010.

[25] H. Çakmak, J. Urbain, and T. Dutoit, "The AV-LASYN database : A synchronous corpus of audio and 3d facial marker data for audio-visual

laughter synthesis," in *Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14)*, 2014.
[26] K. El Haddad, H. Cakmak, S. Dupont, and T. Dutoit, "An HMM Approach for Synthesizing Amused Speech with a Controllable Intensity of Smile," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, UAE, 7-10 December 2015.
[27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
[28] S. Haq, P. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*, Sept. 2008.
[29] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven, "Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, Jun. 2013, pp. 277–282.
[30] W. Ruch and P. Ekman, "The expressive pattern of laughter," *Emotion, qualia, and consciousness*, pp. 426–443, 2001.
[31] M. G. and C. W., "The relationship between laughter intensity and perceived humour," in *The 4th international workshop on laughter and other non-verbal vocalizations in speech*, 2015.
[32] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press, 2014.
[33] G. Bradski, *Dr. Dobb's Journal of Software Tools*.
[34] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1953039
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[36] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[37] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv e-prints*, Dec. 2014.
[38] P. J. Glenn, *Laughter in interaction*. Cambridge University Press, Cambridge, 2003.
[39] W. Chafe, *The Importance of not being earnest. The feeling behind laughter and humor.*, paperback 2009 ed., ser. Consciousness & Emotion Book Series. Amsterdam, The Nederlands: John Benjamins Publishing Company, 2007, vol. 3.
[40] M. S. Edmonson, "Notes on laughter," *Anthropological linguistics*, pp. 23–34, 1987.
[41] W. Ruch, "Exhilaration and humor," *Handbook of emotions*, vol. 1, pp. 605–616, 1993.
[42] J. Urbain, H. Çakmak, A. Charlier, M. Denti, T. Dutoit, and S. Dupont, "Arousal-driven synthesis of laughter," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 273–284, 2014.
[43] I. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications, 1996.
[44] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, Oregon Health & Science University, 2001.
[45] T. Hueber, E.-L. Benaroya, B. Denby, and G. Chollet, "Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface." in *INTERSPEECH*, 2011, pp. 593–596.
[46] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, 2006.
[47] K. El Haddad, S. Dupont, N. d'Alessandro, and T. Dutoit, "An HMM-based speech-smile synthesis system: An approach for amusement synthesis," in *International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Ljubljana, Slovenia, 4-8 May 2015.
[48] S. W. Group *et al.*, "Speech signal processing toolkit (sptk)," *h ttp://sp-tk. sourceforge. net*, 2009.
[49] P. K. Muthukumar and A. W. Black, "A deep learning approach to data-driven parameterizations for statistical parametric speech synthesis," *CoRR*, vol. abs/1409.8558, 2014. [Online]. Available: http://arxiv.org/abs/1409.8558

Kevin El Haddad holds a degree in Microsystems and Embedded systems. He is currently pursuing a PhD in the TCTS lab of the University of Mons (Belgium) under the supervision of Prof. Thierry Dutoit and Dr. Stéphane Dupont. His research interests include affect signals multimodal synthesis and recognition, affective computing and machine learning applications for Human Computer Interactions.



Hüseyin Çakmak holds a double degree in Aeronautics from the Higher Institute of Aeronautics and Space (ISAE) and in Electrical Engineering from the Polytechnic Faculty of Mons (FPMS). He is currently PhD student under a FRIA grant. His research interests are audio and visual synthesis and more specifically audiovisual laughter synthesis using machine learning. He is now a post-doctoral researcher at the signal processing lab of the University of Mons.

Marwan Doumit - iCES corporation, is a software engineer since 2012 currently working for iCES. Marwan holds a computer science degree from NDU (Notre Dame University Lebanon) since 2011 and currently pursuing his masters degree in Computer Science from GMU (George Mason University Virginia USA).

Gueorgui Pironkov holds an Electrical Engineering degree from ISEN (Institut Suprieur d'lectronique et du Numrique) in Lille (France) since 2013. He is currently pursuing a Ph.D degree at the TCTS lab of the University of Mons (Belgium). his research interest include speech recognition and synthesis using deep neural network, multi-task learning and speaker identification.

Uğur Ayvaz got his Bachelor's degree at Mugla Sitki Kocman University (MSKU) and graduated from Computer Engineering Department in 2014. He began his Master's degree at Karadeniz Technical University in 2015. After he had started to work as a research assistant at MSKU, he continued to his Master's degree at Information Systems Engineering department of this university. He is interested in data mining, image and signal processing and machine learning. He aimed to improve his skills in the field of Human-Computer Interaction.

I Probe, Therefore I Am: Designing a Virtual Journalist with Human Emotions

Kevin K. Bowden, *University of California, Santa Cruz,* Tommy Nilsson, *University of Nottingham,* Christine P. Spencer, *Queen's University Belfast,* Kübra Cengiz, *Istanbul Technical University,* Alexandru Ghitulescu, *University of Nottingham* and Jelte B. van Waterschoot, *University of Twente*

***Abstract*—By utilizing different communication channels, such as verbal language, gestures or facial expressions, virtually embodied interactive humans hold a unique potential to bridge the gap between human-computer interaction and actual interhuman communication. The use of virtual humans is consequently becoming increasingly popular in a wide range of areas where such a natural communication might be beneficial, including entertainment, education, mental health research and beyond. Behind this development lies a series of technological advances in a multitude of disciplines, most notably natural language processing, computer vision, and speech synthesis. In this paper we discuss a Virtual Human Journalist, a project employing a number of novel solutions from these disciplines with the goal to demonstrate their viability by producing a humanoid conversational agent capable of naturally eliciting and reacting to information from a human user. A set of qualitative and quantitative evaluation sessions demonstrated the technical feasibility of the system whilst uncovering a number of deficits in its capacity to engage users in a way that would be perceived as natural and emotionally engaging. We argue that naturalness should not always be seen as a desirable goal and suggest that deliberately suppressing the naturalness of virtual human interactions, such as by altering its personality cues, might in some cases yield more desirable results.**

***Index Terms*—Embodied virtual agents, human-computer interaction, anthropomorphic agents, empathy.**

I. INTRODUCTION

THE principal goal of the EU's ARIA-VALUSPA[1] project is to develop a framework for the production of virtual humans capable of engaging in a multi-modal speech-based social interaction with a user in a variety of situations.

In comparison with text-based interactions, speech-based interactions are considered to have a number of well-documented advantages. It has, for instance, been established that direct, first-person speech increases the memorability of a message [1], [2]. Moreover, research also shows that dialogue is more persuasive than monolog [3]. To achieve a natural conversational flow, it is necessary for virtual humans to be able to sustain an interaction whilst reacting appropriately to both the verbal and non-verbal behavioral cues of the user. This demands a range of diverse technological solutions to be in place and operate in unison.

For this purpose, the ARIA-VALUSPA project brings together an international consortium of academic and industrial institutions, including CereProc[2] specialized in the development of text-to-speech (TTS) solutions, a research team from Paris Telecom focusing on the development of virtually embodied human characters and a number of other academic research laboratories across the UK, France, Germany and the Netherlands. In other words, the project builds on a broad and interdisciplinary research area, spanning a multitude of fields, including psychology, linguistics, artificial intelligence and computer graphics.

By summer 2016 the project had progressed to a stage where a prototype facilitating a relatively successful interaction between the user and a virtual human could be produced. To demonstrate the viability of the technology, a spin-off project called Virtual Human Journalist was carried out during the 2016 eNTERFACE workshop. A small team of international students were given access to the technical solutions developed by the ARIA-VALUSPA consortium and tasked to produce a virtual human journalist capable of interviewing and eliciting information from a user. This information was to be stored, processed, and eventually relayed back to the user in a coherent, ostensibly authentic and contextually-appropriate way. In other words, the goal was not simply to develop a working technical solution, but also to produce an experience that would be perceived by users as natural and believable through the demonstration of responsiveness.

Over the course of four weeks, the team developed a dialogue management system centered around a set of probing questions designed to elicit a maximum amount of relevant information from the user. Informal qualitative evaluations of the agent's conversational performance were continuously carried out in parallel to the development of the dialogue system in order to assess its performance in a real world

K. K. Bowden is with the Natural Language and Dialog Systems Lab, University of California, Santa Cruz, USA. Email: kkbowden@ucsc.edu.

T. Nilsson is with The Mixed Reality Laboratory, University of Nottingham, UK, and The National Institute of Mental Health, Klecany, Czech Republic. Email: psxtn2@nottingham.ac.uk.

C. Spencer is with the Social Interactions Lab, Queen's University Belfast, E-mail: cspencer03@qub.ac.uk

K. Cengiz is with the Istanbul Technical University, Turkey. Email: ce.kubra@gmail.com.

A. Ghitulescu is with the Computer Vision Laboratory, University of Nottingham, UK. Email: alex.ghitulescu@gmail.com.

J. B. van Waterschoot is with the Human-Media Interaction Lab, University of Twente, Netherlands. Email: jbvanwaterschoot@utwente.nl.

Manuscript received Month day, year; revised Month day, year.

[1]ARIA-VALUSPA is short for Artificial Retrieval of Information Assistants – Virtual Agents with Linguistic Understanding, Social skills, and Personalised Aspects. Full detail of the project can be found at http://aria-agent.eu/

[2]CereProc, headquartered in Edinburgh, UK, is a leading developer of text-to-speech solutions. Full company overview can be found at https://www.cereproc.com/en/

setting. Besides assessing the perceived quality of the virtual journalist's verbal and non-verbal behavior, a set of quantitative evaluations were likewise aimed at assessing the degree of user emotional engagement and the likability of the agent induced by the system. The overall aim of the workshop was therefore to design and test the efficacy of the system to elicit information from the user whilst maintaining user interest and emotional engagement. This was evaluated by the agent's ability to sustain at least a five minute interaction with the user by using a number of information-eliciting strategies. User interest and emotional engagement were assessed via a post-interaction questionnaire.

A virtual human journalist possessing the ability to successfully elicit information in a sustained and naturally flowing interaction whilst appropriately addressing the user's emotional state may offer multiple valuable future applications. For example, the development of an empathic embodied agent which demonstrates an interest in learning more about the user could offer social companionship to elderly users, or members of vulnerable social groups. In terms of other practical applications, for example, such an agent could reduce labor-intensive human activities by successfully eliciting and storing user's information into a database via a social interaction conducted by telephone.

The following section of this paper provides a brief overview of key theoretical concepts relevant to our work. The third section describes the system design of the virtual human journalist. The fourth section describes the evaluation of the agent, containing two small pilot studies for data collection and testing. In the fifth section we describe some of the limitations of our project. In the sixth section we describe alternative means for improving our system design. Finally, in the seventh section we make concluding remarks about developing an engaging embodied virtual agent for information extraction.

II. Background

A critical factor in interpersonal interactions and relationships is empathy and emotional responsiveness, [4] due to its ability to help individuals to recognize, understand, predict and respond appropriately to the behavior of others [5]. If virtual humans are to achieve interaction quality which is similar, if not equal to, human-human communication, developing convincing simulation of empathic abilities will be crucial. Haase and Tepper [6] suggest that both verbal and non-verbal cues play an important role in the complex multi-channel process of establishing a sense of emotional responsiveness.

Believable empathic cues in virtual humans can consequently come in various forms. In a multi-modal interface, a user might for example perform gestures, such as pointing at literal or figurative objects whilst speaking. It is critical for virtual humans to be able to accurately recognize and "understand" the emotional relevance of such gestures in a particular social context, as well as possessing its own contextually-appropriate gesturing abilities.

Other crucial non-verbal cues include facial expressions. Of particular relevance are the attempts of psychologists to develop facial expression classification systems [7], [8]. For instance, Phillip et al. identified 135 emotion names, which the authors clustered hierarchically [8]. These taxonomies provide us with a broad basis for the mapping of emotion to verbal language, and in turn the opportunity to accompany it with convincing non-verbal signals. The viability of such emotional models has been demonstrated by their successful implementation in video games. [9]

Similarly, various back-channeling signals can indicate to an interaction partner that one is emotionally engaged in an interaction. For example, back-channeling behaviors are thought to demonstrate engaged listenership [10], whilst increased behavioral mimicry can signal greater rapport and an increased desire for affiliation [11]. It therefore follows that, if human-computer interactions are to effectively simulate human-human communication, the agent must display a level of emotional responsiveness and full repertoire of verbal and non-verbal behaviors to be an engaging interaction partner [12].

We centered the virtual human journalist system around a specific knowledge domain from which it can retrieve information pertaining to a particular subject area. The virtual human is also capable of prompting the user to request more information about a pre-defined subject, by asking questions such as "Would you like me to tell you about *X*?", where *X* is an example taken from a list of pre-defined subjects. However, the agent's conversational abilities are currently limited by the constraints of the domain. Given that the goal is to engage the user in a sustained and smooth-flowing interaction, this limitation constitutes a problem. The current conversational dynamic is rather one-sided, and subsequently, less naturalistic and representative of human-human communication. It thus appeared that in order to progress from simply an information-retrieval assistant to a plausible and engaging conversational partner, it would be necessary for the virtual human to demonstrate a more well-defined knowledge base. Moreover, this knowledge base ought also be dynamically expandable with additional information retrieved about the user through a series of probing questions. It was therefore seen as desirable to enhance the existing capabilities of the virtual human by integrating various information-eliciting strategies into its design. Such information-eliciting strategies could involve a range of elements making interaction with the virtual human feel more natural and engaging, including personality cues and simulation of empathy. In other words, it was seen as imperative to design a virtual human in a way that would allow it to engage the user and maintain user interest by demonstrating responsiveness to the user's emotional state. We hypothesize that improving the information gathering in this manner should in turn help expand the knowledge base so that more information could be processed and ultimately relayed back to the user. We predict these abilities will demonstrate a sense of higher cognition and a more sophisticated level of understanding.

III. System Design

Here we describe the implementation of the virtual journalist. We start by elaborating on existing components and how we extended the virtual agent from the ARIA-VALUSPA project to become a virtual journalist. We will close this section with an example of a dialogue and how the pipeline processes the user input in this example. "Alice" was the agent selected for use as the interviewer, adopted from the ARIA-VALUSPA project[3]. The agent understands spoken natural language and non-verbal behavior and uses both verbal and non-verbal behavior to converse with users. Alice's visual appearance is displayed in Figure 3.

A. Preprocessing

The input is first preprocessed by implementing components from the Stanford Natural Language Parser (NLP) toolkit [13]. Specifically we used the dependency parser and subsequently the part of speech tagger to determine the structure of the sentence. We identify terms as important based on their part of speech and syntactic role within a sentence. We create nodes in a hand crafted knowledge base for important nodes within the sentence such as the named entities, proper nouns, and the subject. Additionally, we track the attributes which modify a node, possessions a node claims ownership over, and the other names a node can be referred to as by using light weight handcrafted anaphoric resolution. We also attempt to identify nodes which appear to be related to each other based on frequent co-occurrence in the same context.

B. Dialogue Manager

Flipper served as the dialogue manager (DM), structuring the dialogue dynamically between the agent and the user, based on an information-state approach to dialogues [14] [15]. Flipper does this by keeping track of the verbal (e.g. subject, proper nouns) and non-verbal (e.g. emotion) input from the user in the IS [14]. Based on certain keywords and non-verbal user behavior defined in Flipper's templates, the agent would formulate its intent, i.e. how it wants to respond to the user. Every such possible response was written manually using the Functional Markup Language (FML), each defining the appropriate semantic units for the specific communicative intent [16] [17]. The advantage of using FML is that the dialogue manager only has to care about selecting an agent intent, while the behavior planner fills in the blanks for how exactly to do perform the intent. For example in the DM you can define to say something angrily, and the behavior planner will decide if this is expressed via facial expressions or changing the pitch. The workings of the behavior planner are worked out in the next section.

Hey, my name is Alice, what's your name?
Nice to meet you {name}. So what do you do for fun in your spare time?
So what do you specifically like about {X}?
So, you spoke a bit about {Y}. Why don't you tell me more about {Y}?
Nice, OK. Maybe we can chat about something else now. Do you have any pet?

Table I: Example of conversational questions used by Alice

[3]https://github.com/ARIA-VALUSPA/ARIA-System

We formulated questions of which some were emphatically phrased, based on findings of our first pilot study. Questions were designed to elicit usable nouns of the user, whilst being general enough to be applicable to most contexts, such as "*What do you do for a living?*". Each subsequent response would then be dependent on the user's input. For example, if the user provided a long, informative answer, the agent could ask a few follow-up questions until the topic was exhausted. Conversely, if the user did not provide an adequate amount of information about a topic (*X*), the agent could ask a question prompting more elaboration, such as "*I see. I'd be very interested in hearing more about X. Could you tell me a little bit more about that?*". If the user's response was too complex, the agent then asked the user to rephrase the information.

Finally, if the conversation began to run dry, or the topic became exhausted, the agent asked a question prompting a new topic. The speech content was additionally categorized by emotional valence, in order for the agent to adequately address the emotional content of the user's response with appropriately timed facial expressions and gestures for displaying empathy. Each state was also linked to a pool of responses such that the system can trigger the correct state while also avoiding exact repetitions. Examples of probing questions are included in Table I.

In this stage we augment Flipper's ability to pick an appropriate response using the knowledge base we built in the preprocessing phase. We use Equation 1, to determine how relevant a node is to the current context. Specifically, we account for three node facets. We are interested in the frequency of a node - the number of mentions within the conversation, the time since it's been last mentioned, and the preference of a node. Here we establish the preference of a node by examining the sentiment of the content for which it's been applied. Preference is represented on a scale of 0-1 where our system sets the default preference of a given node to be neutral with a score of .5. We leave the robustness of this feature for future work.

$$score(u) = (freq(u) - \frac{timeSinceLast(u)}{1000})pref(u) \quad (1)$$

Flipper determines if there is a state which we can transition to that leverages our knowledge with a high scoring utterance. If we find such a state, we perform the necessary substitutions to make a valid and relevant response. If we are unable to leverage any of our information, Flipper uses its default keyword matching to determine the next state or to switch to a new topic generically. This default keyword matching also takes into account synonyms from a handcrafted list via the original ARIA-VALUSPA project. The DM is also responsible for keeping the conversation moving, if the user hasn't responded in a set amount of time we'll prompt the user to continue the conversation.

C. Output

After the agent response was selected, Alice executed the behavior based on FML [18]. The FML could contain parameters such as where to apply a pitch accent or which emotion

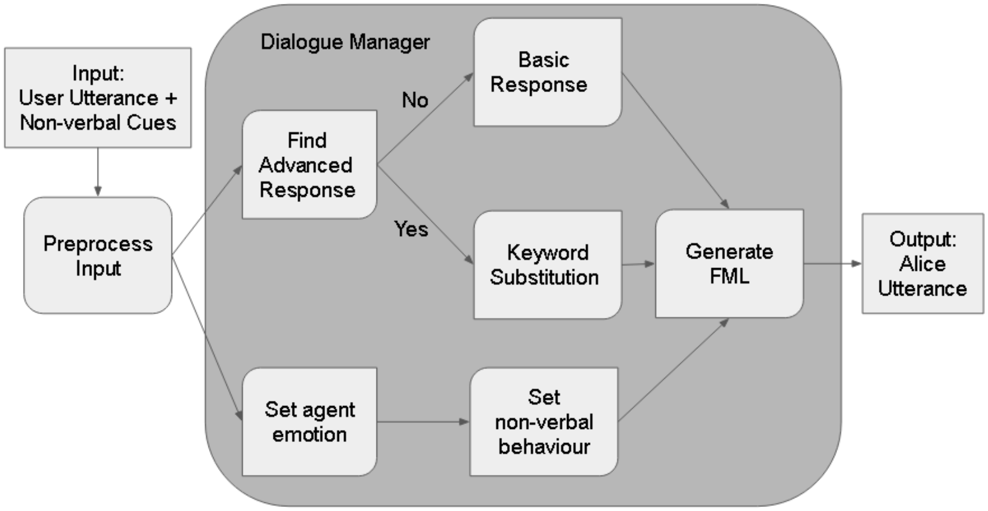


Figure 1: Our general system pipeline

the agent has to show. These parameters can be loaded from the IS of the DM (e.g. deciding to mirror the emotion of the user) or be fixed in the FML-files (e.g. a question that always displays 'surprise' behavior). An example conversation with Alice is included in Figure 2.

Alice: OK, maybe we can talk about something else now. If you won the lottery tomorrow, what would be the first thing that you would do?
User: The first thing I would do? Wow that is a difficult question. I would travel around I think to visit new places, new countries.
Alice: Where would you go?
User: To visit Asia, I think, or Australia.
Alice: Nice! I have never been to Australia before. Can you tell me about it?
User: I've never been there as well.
Alice: I see. Maybe we can both go there one day.
User: Yeah, sure!

Figure 2: An example conversation with Alice

IV. Evaluation of the Agent

We did a two-step evaluation of our system, first on a user to wizard basis, second on a user to agent basis.

A. Wizard of Oz Evaluation

In order to get an early indication of its real world applicability (e.g. what does an interviewer do and how does empathy play a role), an informal evaluation of the agent begun whilst still under development, using a semi-functional version that was operated via Wizard of Oz set up [19], [20]. The interactions were recorded and transcribed for analysis. A second goal of our wizard was to generate a list of generic questions which both elicited good follow-up conversation and were generic enough to apply in different contexts.

Metric	$Mean_1$	$Mean_2$
Liked Agent	2.75	3
Likes Agents	3.38	3.72
Engaging	2.75	3.14
Enjoyment	2.75	3.14
Vocabulary	2.63	2.72
Naturalness	1.88	2.15
Clarity	2.63	2.86
Background Knowledge	3.63	3.43
Background Experience	2.75	2.57
Agents Behaviors	2.63	2.72
Conversational Abilities	3.13	3.01
Empathic Abilities	3.13	2.86

Table II: These are the heuristics our system uses to determine what sort of variation to insert.

Figure 3: Alice - the virtual human journalist

B. Agent Evaluation

Using this refined list of questions we performed a series of interviews with 8 participants (5 male, 3 female). All agent responses were triggered by a human wizard hidden in a separate room with a list of possible responses, manually determining which was best to use. The participants interacted with the wizard for approximately 5 minutes before completing a 16 question post-test questionnaire. The aim of the evaluation survey was to quantitatively assess each user's enjoyment and engagement during the interaction, their perceptions of the naturalness of the interaction, as well as their level of previous knowledge and experience of virtual agents. Users' impressions of the agents likability, depth of vocabulary, quality of non-verbal expression (gestures, facial expressions, timing), clarity of communication, overall conversational abilities as well as her empathic abilities were also assessed. Each metric was defined on the survey and evaluated on a scale of 0-5 ranging from very low to very high, the mean scores $(mean_1)$ for each metric are included in Table II. In our evaluation one of the participants cited their consistently low scores with system latency due to internet problems which continually interrupted the session, we've included the mean scores $(mean_2)$ without this participant as well. A qualitative outlet was also provided in order to gather user feedback regarding ideas for the future applications of the agent.

Analysis of participant feedback indicated that participants were generally able to adequately communicate with the agent and answer the questions posed. It was found that the agent evoked a range of responses. The journalist was generally perceived as moderately engaging, with 7 participants reporting the interaction to have been at least moderately to highly enjoyable. However, most participants scored the interaction low on naturalness. Both the verbal and non-verbal communication of the agent received mixed ratings. Only half of the participants found the agent's conversational abilities to be good, whilst the agent's clarity of communication scores particularly varied across participants. The empathy scores reflected that the agent's empathic performance was experienced as being relatively poor. 6 out of 8 participants reported liking virtual agents to some degree, while two were more negative. In terms of suggestions for the potential future applications of the agent, most users suggested that an agent like Alice would be well-suited to the domain of virtual tutoring, or operating as a form of virtual receptionist or social companion, the latter being closest to the planned use of the agent from the ARIA-VALUSPA project's perspective.

V. Discussion

The aim of this project was to design and test the efficacy of a virtual agent's ability to elicit information from the user whilst maintaining user interest and emotional engagement. The feedback garnered during the user evaluations showed that the virtual human journalist performed moderately well during the piloting stage, particularly regarding the duration of sustained conversation held between the agent and user. Most interactions lasted the entirety of the 5 minute allocation without petering out naturally, and were brought to a close by the agent. Therefore our dialogue strategy appears to have been successful in maintaining user engagement.

Scores such as naturalness may have been adversely affected by signaling difficulties which were encountered during testing. The agent did not always articulate words correctly, and some sentences were incoherent, or difficult to hear at times. Additionally some users reported a notable delay in the systems response time, we attribute this to our blackbox TTS/STT (speech-to-text) interface. It should be noted that, as all of the participants were recruited opportunistically during the workshop for the pilot testing stage, all consequently possessed at least some degree of experience and knowledge about virtual agents. While it's evident that a number of shortcomings generally stemmed from technical limitations, the agent's empathic performance was notably poor due to the limited semantic and emotional content of the expressions and gestures available for use. Specifically, user feedback also revealed that the agent's prosody and pronunciation was poor and often contextually-inappropriate.

Emotional responsiveness is an imperative phenomenon associated with human nature [21] and the ability to represent functional social bonds [22]. An agent which attempts to depict a human persona is likely to cause the user to form subconscious expectations as to the agents inherent abilities with respect to empathic understanding. An agent which attempts to depict a human persona is likely to cause the user to form subconscious expectations as to the agents inherent abilities with respect to empathic understanding. In these cases users frequently engage in these over-learned social behaviors, such as politeness and reciprocity, when engaging with traditional computers [23], traits which a naive agent cannot properly leverage. Therefore, an agent which puts on this human-like persona but fails to be emotional responsive would immediately be subconsciously recognized as deficient - forming an off-putting environment for the user. Conversely, an agent presented with fewer anthropomorphic cues, subsequently initiating lower user expectations of human-like behavior, could be more easily forgiven for failing to do so.

Having less anthropomorphic cues would lead to a need for the agent to engage the user in a different way, perhaps by demonstrating other qualities of human nature, such as agency, individuality, cognitive openness and depth of mind [21], which may be more convincingly simulated compared to empathy. Whilst an agent depicted as less human could be penalized less for failing to display appropriate emotional cues, such an agent could theoretically be rewarded more for simulating higher cognition, and the ability to engage in deeper levels of abstract thought. Future work may be aimed at testing whether a less humanoid "philosopher-type" character could be capable of more successfully engaging the user. Such an agent might for instance surprise the user by appearing to be in the midst of an existential crisis, and questioning the meaning of her existence. The agent could be presented to the user in a number forms, varying in terms of politeness in addition to positive and negative personality features. An important point may be for the agent to openly admit to the user to being currently in a learning phase, but also for the agent to demonstrate a genuine eagerness and desire to learn more about what it is to be human.

VI. Future Work

It was found that Alice failed to engage users emotionally. There are a number of challenges involved in the development of agents that can convincingly simulate empathy and demonstrate emotional responsiveness. Indeed, Ochs, Pelachaud and Sadek [24] found that users experience agents who respond in an incongruous emotional manner to be more off-putting than agents which do not respond emotionally at all. Although Alice was presented visually to users as being human, it was clear to users that she lacked a deeper understanding of the information presented to her, evident by her lack of emotional responsiveness as a consequence of the limitations of the system.

The likability scores of the agent reflected that the character Alice was generally liked, but anecdotal feedback from the users who had interacted with her indicated that perhaps her overt politeness may have made her a slightly generic and unmemorable character. Schneiderman [25] argues that an overly friendly and "humane" agent should be avoided, as this might make users conclude that the system is more intelligent than it actually is and trust it too much. Particularly experienced users may even experience simulated friendliness as annoying and misleading. Microsoft Bob is an example of such an overtly friendly natural language system that was rejected by the general public [26].

A number of solutions relying on limited anthropomorphic features is already finding its way to the mainstream consumer. Perhaps most notable is the rising tide of intelligent personal assistants, such as Siri. Unlike other voice-controlled services passively awaiting the user command, Siri has been designed to communicate proactively, and even joke with the user, thus strengthening a sense of human personality. However, Siri lacks any form of visual embodied representation, thus limiting a sense of human presence. Similar solutions are currently under development for domestic IoT systems [27].

A further point of interest would be to examine whether users would be more likely to engage with an embodied virtual agent with more notable or memorable personality features. Future work may be aimed at determining the agent's levels of politeness and displays of overt engagement at which the user retains interest. In other words, it might be useful to determine at what threshold for which politeness and displays of overt engagement does an agent begin to lose user interest. Other future work could examine how users respond to an anti-polite character who is opinionated, argumentative and sometimes rude, and from which point does the user become less amused and engaged and start to find these personality features irritating or disengaging [28]. Provisional work was conducted on this in the beginnings of development of a character called Ursula, who was anecdotally found to be more amusing to interact with than Alice.

VII. Conclusion

Promising first steps were made toward the development of an embodied virtual agent. The agent was capable of utilizing a number of information-eliciting strategies in order to achieve a sustained interaction with a user. We argued that emotional engagement can be improved by enhancing the empathic capabilities of the agent through the integration of automatic emotion recognition and social-signaling software. Additionally, machine learning techniques can automate the process of increasing speech template density. Finally, we hypothesize that users may find a less anthropomorphic agent which exhibits more memorable or unexpected personality features to be more interesting and engaging.

Acknowledgment

The authors would like to thank Professor Michel Valstar and his ARIA-VALUSPA team for granting us access to their development tools. We would also like to extend our thanks to the eNTERFACE workshop organizers for making this project possible.

References

[1] D. Tannen, "Talking voices : repetition, dialogue, and imagery in conversational discourse," *CUP*, 1989.
[2] D. Schiffrin, "Tense variation in narrative," *Language*, 1981.
[3] P. Piwek, H. Hernault, H. Prendinger, and M. Ishizuka., "T2d: Generating dialogues between virtual agents automatically from text.," *In Intelligent Virtual Agents*, 2007.
[4] J. Decety and C. Lamm., "Human empathy through the lens of social neuroscience," *The Scientific World Journal*, 2006.
[5] D. V. Frederique and S. Tania, "The empathic brain: how, when and why?," *Trends in cognitive sciences*, 2006.
[6] R. F. Haase and D. T. Tepper., "Nonverbal components of empathic communication," *Journal of Counseling Psychology*, 1972.
[7] E. Fast, B. Chen, and M. S. Bernstein., "Understanding topic signals in largescale text..," in *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16),*, 2016.
[8] D. K. Phillip R. Shaver, Judith Schwartz and C. O?Connor., "Emotion knowledge: Further exploration of a prototype approach," *Journal of Personality and Social Psychology*, 1987.
[9] A. Chowanda, P. Blanchfield, M. Flintham, and M. Valstar., "Computational models of emotion, personality, and social relationships for interactions in games," in *In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 2016.
[10] L. Kathrin, "Back-channelling: The use of yeah and mm to portray engaged listenership.," *Griffiths Working Papers in Pragmatics and Intercultural Communication*, 2011.
[11] L. Jessica and C. Tanya, L., "Using nonconscious behavioral mimicry to create affiliation and rapport.," *Psychological science*, 2003.
[12] G. McKeown, "Turing's menagerie: Talking lions, virtual bats, electric sheep and analogical peacocks: Common ground and common interest are necessary components of engagement.," *In Affective Computing and Intelligent Interaction*, 2015.
[13] K. Toutanova, D. Klein, and C. Manning, "Feature-rich part-of-speech tagging with a cyclic dependency network," pp. 252–259, 2003.
[14] M. ter Maat and D. Heylen, "Flipper: An information state component for spoken dialogue systems," in *International Workshop on Intelligent Virtual Agents*, pp. 470–472, Springer, 2011.
[15] D. R. Traum and S. Larsson, "The information state approach to dialogue management," in *Current and new directions in discourse and dialogue*, pp. 325–353, Springer, 2003.
[16] D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson, "The next step towards a function markup language," in *International Workshop on Intelligent Virtual Agents*, pp. 270–280, Springer, 2008.
[17] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, *et al.*, "Towards an iso standard for dialogue act annotation," in *Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.
[18] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive eca system," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 1399–1400, International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[19] B. Hanington and M. Bella, *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Publishers, 2012.
[20] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies: why and how," in *In Proceedings of the 1st international conference on Intelligent user interfaces*, 1993.
[21] N. Haslam., "Dehumanization: An integrative review," *Personality and social psychology review*, 2006.
[22] A. Rumble, P. V. Lange, , and C. D. Parks, "The benefits of empathy: When empathy may sustain cooperation in social dilemmas," *European Journal Of Social Psychology*, 2010.
[23] C. Nass and Y. Moon., "Machines and mindlessness: Social responses to computers," *Journal of social issues*, 2000.
[24] M. Ochs, C. Pelachaud, and D. Sadek., "An empathic virtual dialog agent to improve human-machine interaction," *In Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, 2008.
[25] B. Schneiderman, *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education, 2010.
[26] "Microsoft Bob." Microsoft Bob.
[27] Ivee, Talking to Your Home is No Longer Sci Fi.
[28] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

Proceedings of eNTERFACE'16, Enschede, July 18th – August 12th, 2016

Development of low-cost portable hand exoskeleton for assistive and rehabilitation purposes

Matteo Bianchi, *Student Member, IEEE,* Tobias Bützer, *Student Member, IEEE,* Stefano Laszlo Capitani, Arianna Cremoni, Francesco Fanelli, *Student Member, IEEE,* Nicola Secciani, Matteo Venturi, Alessandro Ridolfi, *Member, IEEE,* Federica Vannetti, and Benedetto Allotta, *Member, IEEE*

***Abstract*—The design of an aid for the hand function based on exoskeleton technologies for patients who have lost or injured hand skills, e.g. because of neuromuscular or aging diseases, is one of the most influential challenge in modern robotics to assure them an independent and healthy life.**

This research activity is focused on the design and development of a low-cost Hand Exoskeleton System (HES) for supporting patients affected by hand disabilities during the Activities of Daily Living (ADLs). The device can be also used during the rehabilitative sessions to better recovery the dexterity of the user's hand.

This paper presents a compact design concept for a portable hand exoskeleton.

This prototype has been developed thanks to the collaboration between the Department of Industrial Engineering (DIEF) of the University of Florence, and the Rehabilitation Engineering Laboratory of the ETH, Zürich, during the eNTERFACE16 Workshop, hosted by the University of Twente.

***Index Terms*—Wearable Robotics, Portable Devices, Rehabilitation, Human-oriented Approaches**

I. Introduction

THE demand for rehabilitation therapy has been increasing rapidly in the last years. Two aspects are contributing significantly to this situation. The former is that aging is a global problem and there will be an increasing number of elderly people who need resources for rehabilitative therapy. The latter is that epidemiologically, not only as a consequence of aging, more people will suffer from neurological and musculoskeletal disease such as stroke, Parkinson's Disease and degenerative arthritis [1].

From the viewpoint of rehabilitation, it is crucial for the patient to perform intensive and continuous therapeutic tasks for a successful rehabilitation.

Robotics systems allow for prolonged and high-intensity rehabilitation treatments, with a reduction of costs and burden for the therapists. This devices are also able to evaluate the patients' progresses by measuring physical parameters and to replicate a given protocol always in the same conditions. Both of this operations are difficult with manual therapy. Since such devices [2] are designed to be used during rehabilitative sessions, their portability requirement is not mandatory, as they are used in hospitals, rehabilitation centers or at home in a specific location dedicated to their utilization.

Wearability and portability requirements become significant, instead, in those cases where the hand exoskeletons have to be used in order to assist the user in ADLs. In fact, hand functions may not be totally replaced even after an intense rehabilitation process (e.g. up to 66% of post-stroke patients have not regained the dexterity of the damaged arm after 6 months from the stroke [3]). In these cases, such devices can be used to assist the hand performance by amplifying the hand gripping force [4] or automating the motion[5].

In healthcare, there is a real need for novel devices which are modular, cost-effective, easy to use, and extremely reliable. These tools need to have flexibility to meet various sets of requirements and also societal expectations. Novel methods and tools need also to be scalable and adaptable, so that they can be used by different patients and with different kinds of settings. New solutions should be easy to use and they should be accepted by end-users: elderly/patients, medics, care service providers but also insurance companies and such.

Unfortunately, considering the devices described in literature, only few of them present an outcome of their use in the clinical practice as positive as expected. Owing to strict motor and sensor requirements in terms of mechanism, weight, size and dexterous manipulation capabilities, portable hand exoskeletons for rehabilitation and assistance [6] have not been developed as well as the exoskeleton robots for lower and upper limbs.

For all these reasons, the design of an aid for the hand function based on exoskeleton technologies for patients who have lost or injured hand skills because of neuromuscular or aging diseases, is one of the most influential challenge in modern robotics to assure them an independent and healthy life.

In this paper the researchers of the Mechatronics and Dynamic Modeling Laboratory (MDM Lab) of the Department of Industrial Engineering (DIEF) of the University of Florence have worked together the researchers of the Rehabilitation Engineering Laboratory of the ETH, Zürich, in order to redesign an already built device mainly realized in Acrylonitrile Butadiene Styrene (ABS) structural components, with a cable-driven actuation provided by means of four independent servomotors placed on the back of the hand [7]. That device is characterized by a 1-DOF mechanism which replicates the trajectory of the

M. Bianchi, S.L. Capitani, A. Cremoni, F. Fanelli, N. Secciani, A. Ridolfi and B. Allotta are with the Department of Industrial Engineering (DIEF), University of Florence, Via di Santa Marta 3, 50139, Florence, Italy. e-mail: matteo.bianchi@unifi.it.

T. Bützer is with the Rehabilitation Engineering Laboratory, ETH Zürich,, 8092 Zürich, Switzerland.

F. Vannetti is with the Don Carlo Gnocchi Rehabilitation Center, Florence, Italy.

B. Allotta is with the Department of Industrial Engineering (DIEF), University of Florence and with the Don Carlo Gnocchi Rehabilitation Center.

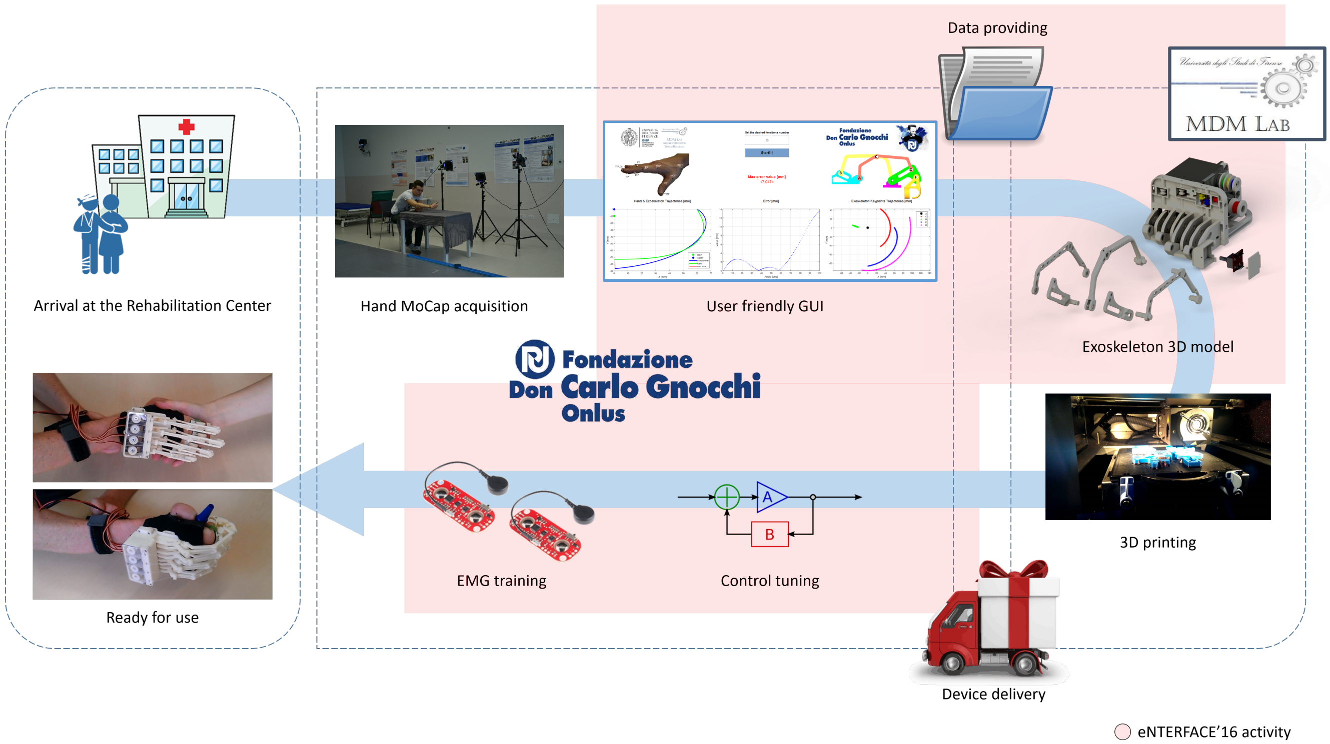


Fig. 1. Proposed scenario

second phalanx of the finger.

In this context, eNTERFACE'16 gave the opportunity to investigate and assess the development of a novel system which, starting from the geometry of a hand, leads to the build of a portable hand exoskeleton. Each step, necessary to achieve this goal, are reported in the following Section, illustrating the workflow of the project.

II. General Architecture

The research activity described in the paper aims at developing the already-designed device, mentioned in the last part of Section I, in order to provide the rehabilitation center of real devices which could be used as an active aid for patients. The proposed scenario is reported in Fig. 1.

Referring to Fig 1, the following sections of the paper will explain the main steps that lead the patience to be given his/her most appropriate device. After the arrival at the rehabilitation center, the hand motion of the patient can be studied and the trajectory of all fingers can be acquired by means of a Motion Capture (MoCap) system (Section III). Then, thanks to a specific scaling algorithm developed by the MDM Lab, it is possible to determine the suitable shape of the fingers mechanism which replicate at best the acquired hand trajectories (Section III). In this part of the scenario an user-friendly Graphic User Interface (GUI) has been built to help the clinicians during the acquisition. The GUI suggests the right position of the markers on the hand, lets the clinician to set the inputs iterations for the scaling algorithm (Section III) and provides some outputs (i.e. the max and average error in following the hand trajectory by the mechanism) useful to check the goodness of the results. Once the dimensions of the exoskeleton are determined, data are provided to the MDM Lab, where the exoskeleton is build by means of the 3D printer of the laboratory. The new design of the device is reported in Section IV. Finally, the exoskeleton is sent to the rehabilitation center and, after a brief tuning of the control system (described in Section V) and a short training for the patient, it is ready to be used. In Section VI some results concerning the proposed system are reported and conclusions and further development will be discussed.

Taking in mind the proposed scenario, in Fig. 1, pink squares highlights the activity carried out during eNTERFACE'16 workshop. Since the previous prototype was thought to be used only on a specific patient, a scaling algorithm has been developed to adapt the former mechanism to different users. Making the procedure fully automatic required the development of a GUI as easy to use as possible. This GUI has been integrate in the scaling code and the code itself has been improved to provide directly a 3D CAD model immediately printable. This part of the work represents an important improvement with respect to the previous preliminary work presented in [16].

Finally, the novel EMG triggered actuation strategy, based on the exploitation of only 2 electrodes, has been investigated in this context for the first time.

Unfortunately, during the workshop, 3D printer was available only for PLA and without support material. Since the complex geometry of the exoskeleton parts required support material, a physical tested embodiment of the prototype was not possible.

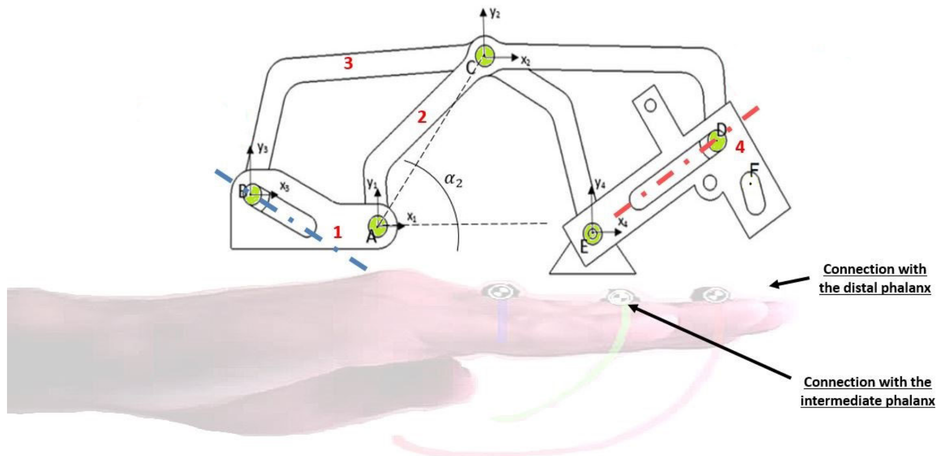


Fig. 2. The mechanism of one finger of the exoskeleton

III. Scaling

Since the anatomical characteristics of the hand considerably vary between different patients both because of different anthrophometries and different pathologies, a procedure to adapt the exoskeleton to generic patients is needed.
An optimization based method is proposed to adapt the mechanism to different users. Each resulting device is specific for the hand anatomy of the patient and is able to closely follow the trajectories of all the long fingers.
The whole procedure can be divided into two steps.
First of all, the motion analysis of the hand kinematics is required in order to obtain the data whose use constitutes the second part of the process. The authors exploited the BTS SMART-Suite Motion Capture System by BTS Bioengineering [13] available at the Don Carlo Gnocchi Foundation Rehabilitation Center, Florence, Italy [14]. This optoelectronic system is made up of 4 infrared cameras with an acquisition rate of 100 Hz and is able to automatically record 3D trajectories of markers, placed on the hand of the patient, by means of stereophotogrammetric methods.
The second step consists in changing the dimensions of the mechanical parts of the exoskeleton mechanism to match the trajectories acquired by the MoCap.
Refer to Fig. 2 which shows the mechanism of one finger of the exoskeleton.
By changing the relative distances between the points A, B, C, D and E, the trajectory of point E, which is the link between the mechanism and the finger, is modified. The objective of this step is to determine such distances so that the exoskeleton follows the natural motion of the fingers. To achieve this goal, the authors resorted to numerical optimization; in particular, the method proposed in [15], which is a Nelder-Mead based optimization algorithm used to solve non convex, non linear constrained problems, has been used. In fact, taking the acquisition data and the kinematics of the mechanism as inputs, the employment of the implemented algorithm provides the customized geometry specific for each user.
The aforementioned steps are explained in details in [16].

IV. Redesign

The portability of the hand exoskeleton is strictly related to its weight and size, while its usability relies on its high stiffness. Because of such specifications in terms of small size (and low weight) and high stiffness, a redesign of the previous version [7] of the exoskeleton is carried out to improve its wearability by reducing its weight without a lowering of its stiffness.
In particular, at first, the number of the servomotors is reduced from four to two (one for the index finger and one for the other three long fingers). Since the mechanisms for each finger have different sizes, the opening and closing velocity is different for each finger. Through the design of a particular pulley with three different diameters (Fig. 3), it is possible to actuate middle, ring and small finger mechanism at the same time with the same motor. *This solution reduced weight by* 59% *with respect to the previous version (*$405g$ *vs* $242g$*).* Then, the dimensions of the servomotors housing, from 67x63.5x84 mm, become 48x66x74 mm.

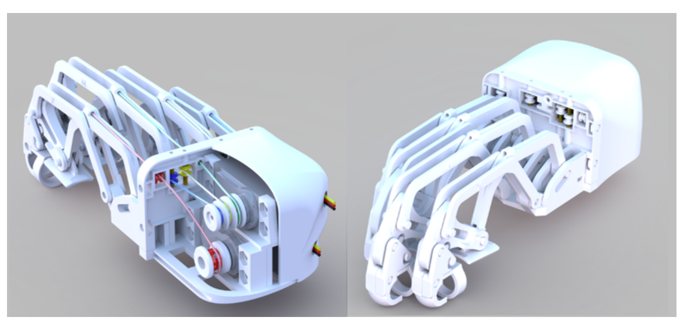


Fig. 3. Hand exoskeleton rendered image

A further weight reduction, with no loss in stiffness, in this new design of the finger mechanism was achieved by means a topology optimization method [12]. This technique has the capability of generating a large number of degrees of freedom available for the design variable settings, so that the optimal design can be obtained without prior knowledge.

Based on the existing kinematic study, an analysis performed with MATLAB® was carried out in order to find the maximum reaction forces on the mechanism during opening and closing gestures. In particular, a force of 10 N is exerted on the finger (point E of Fig. 2) by means of a spring between points D and E of the mechanism (see Fig. 2).
This solution makes possible a passive actuation of the closure gesture, while the opening gesture is actuated by a cable driven mechanism. Such force value is chosen to allow the patient to grab common objects during the Activities of Daily Livings (ADLs). During the opening phase, the cable exerts forces on the finger, but, due to the presence of the spring, the reaction forces on the mechanism are lower. Hence, the topology optimization focused only on the closure gesture.
The optimization of the mechanism is carried out using SolidThinking Inspire®, a structural simulation software included in the Altair suite.
To reach maximum control of the topology optimization process, every part was optimized separately with the software tools applying to them the forces calculated using MATLAB® model. At the end of the optimization process, a 25% weight reduction has been obtained for the finger mechanism, as it is explained in details in Tab. I.
The ring finger has been chosen as case study, because it is the one characterized by maximum mechanical stresses. With minor changes in dimensions due to the hand anatomy, the kinematic chain is repeated for all the fingers. The obtained results can therefore be safely extended to the entire hand. A maximum displacement of 2 mm, for the most elongated bodies is taken as design target value for each component stiffness.
In order to pursue this approach, however, a load analysis is required in order to evaluate the forces in the various joints as the configuration of the mechanism (defined by a single DOF) changes (i.e. motion of the finger).

TABLE I
WEIGHT REDUCTION FOR THE FINGER MECHANISMS AFTER THE TOPOLOGY OPTIMIZATION PROCESS [g]

Finger	Before	After	%
Index	10.54	7.63	27.59
Middle	11.79	9.99	15.24
Ring	11.17	8.68	40.94
Small	9.21	5.44	25.67
Total	42.71	31.74	25.67

A. Kinetostatic Analysis

Basing on the existing kinematic study [7], the kinetostatic analysis was performed in MATLAB® environment by using a 2D model with two redundant multibody methods (Newtonian and Lagrangian approaches [16]). This analysis has been carried out with an isostatic configuration, inserting the force exerted by the aforementioned spring and simulating the real operating conditions (thus removing the only system DOF). This way, a force of 10 N, placed at the extremity of the kinematic chain (point E in Fig.2), is obtained and it represents the force required for grasping.
The trend of joint reactions are evaluated for each pose of the mechanism during the entire Range Of Motion (ROM), corresponding to a ROM of 90° for the joint A of Fig. 2. The 0° angle corresponds to a configuration with the fingers totally extended.
The results obtained with both approaches are perfectly comparable and provide a complete description of the mechanism loads during the finger opening/closing gestures. Inertial effects were omitted in this study, considering the low accelerations. In fact, in reality, finger opening/closing movements are performed with an approximately constant angular speed of $0.6 rad/s$ for the angle α_2 of Fig. 2.

B. Topology Optimization

The original volumes of the parts were considered as a starting point. These values represent the maximum dimensions allowed for the optimized parts and cannot be further enlarged, due to the multiple constraints that need to be considered (e.g. the kinematic of the mechanism, the presence of the finger and other exoskeleton elements). The topological optimization is carried out using the maximize stiffness approach, in line with the design requirements. With these settings, Inspire®searches for the shape solution that maximizes the stiffness using the minimum quantity of material; the obtained result, therefore, has the most efficient shape for the considered application.
To safeguard the kinematic functionalities of the mechanism, the features directly responsible for coupling of the elements (e.g. slots, rotational joints) are left out of the optimization alterable volumes.
A value of 2 mm, consistent with the 3D printing production features, was set as admissible minimum thickness for the optimized volumes. In Fig. 4 the model obtained at the end of the topological optimization is reported.

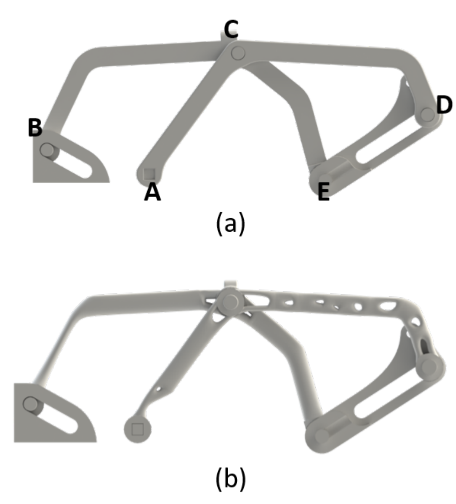


Fig. 4. Hand exoskeleton finger mechanism before (a) and after the topological optimization process

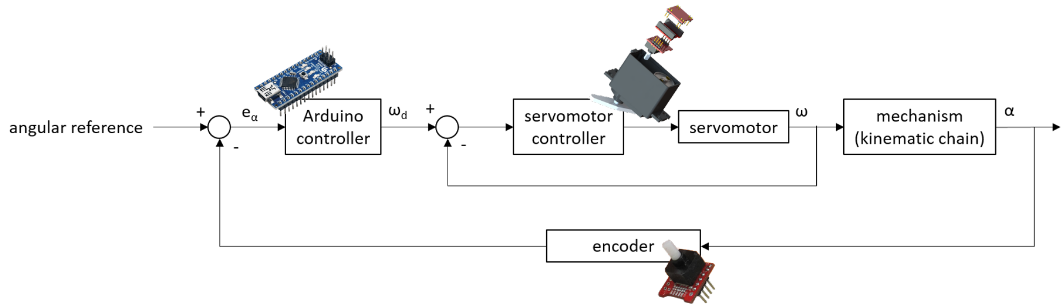


Fig. 5. Closed-loop angular control

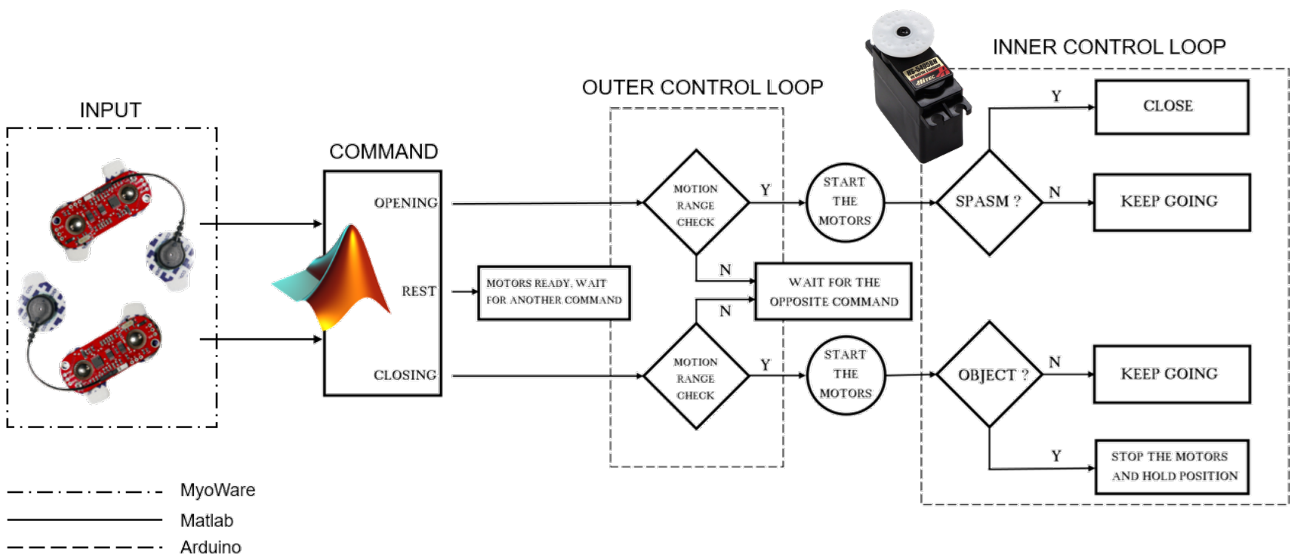


Fig. 6. Control architecture of the system

V. Control

The control system has been designed with a precise goal in mind: keeping its weight, complexity and costs as low as possible. Lightness, cheapness and simplicity are some of the main characteristics that make a device suitable for the application to a large number of people. An EMG-driven control system has been chosen to make as intuitive as possible for patients to use the device. According to all these ideas, the followings have been chosen as parts of the system shown in Fig. 5:

- *micro-controller: Arduino Nano, represented a very good trade-off between performances, simplicity and cheapness. The 16MHz-clock processor was enough powerful to work with signals whose maximum frequency is about 500Hz. The embedded board offered the possibility to directly connect lots of sensors already present on the market - drastically decreasing the complexity of connections - while its elementary programmability makes it easily re-configurable. Moreover, other cases in literature have been found in which Arduino has been successfully used for robotic applications, e.g. [8].*
- *trigger command sensors: two MyoWare Muscle Sensors, placed on the Flexor Digitorum Superficialis and on the Extensor Indicis muscle, respectively, and providing the envelope (amplified, rectified and integrated signal) of the raw EMG signals.*
- *feedback sensors: two 15-bit magnetic encoders, placed on the joint A of the mechanism - see Fig. 2 and Fig. 4 - of the index and little finger, respectively, and providing the value of the angle α_2, which identifies the single DOF of the mechanism.*
- *actuators: two HS-5495BH servo motors - one actuating index and one actuating the other three fingers - placed on the back of the hand, see Fig. 3. According to our low cost concept, many solutions that can be found in the state of the art have been avoided due to their higher costs [9], [10]. Such motors have a maximum torque of* .063 *N*m @6.0 V and* 0.73 *kg*cm @7.4 V with a size of 39.8x19.8x38.0 mm and weight of* 45 *g. The maximum angular speed is* 6.15 *rad/s @6.0 V and* 6.67 *rad/s @7.4 V (performance specifications are reported on the servomotors datasheet). Another important point that has been taken into account lied in the possibility to exploit customized and dedicated Arduino libraries provided by the servo producer itself.*

The control strategy, see Fig. 6, has been thought to start from a command triggered by muscular activity read by the MyoWare sensors at rate of 1024Hz. The signal pair provided by the two muscle sensors is used to feed a MATLAB® built-in Linear Discriminant Analysis (LDA) classifier [11] which provides a real-time classification of three different gestures: opening, closing and resting. Before each use of the exoskeleton, the classifier has been trained with 250 *samples per gesture, with* 10 *seconds training period for each one. Once the trigger is received, Arduino transmits it to the actuation system and starts the control loop. The main control scheme prevents the exoskeleton from finishing the opening or the closing gestures out of the fixed range of motion. This is done using feedback coming from the two magnetic encoders*[1]*. Moreover, two inner control loops were added to take care of the grasping of objects and to indulge the hand motion if an unexpected muscular spasm occurs during the opening phase. Grasping of an object is detected calculating the length of the unrolled cable twice - the first time using the kinematic equations of the mechanism and the second one using the motors speed and the pulleys radius - and then comparing the differential measurement to a set threshold. Muscular spasms are detected, instead, when the instant motor speed falls below a set percentage of the nominal speed. In the first case the motors stop and hold their angular position while in the second case they invert their motion, from opening to closing.*

VI. Results and Conclusions

The final version of the design HES is reported in Fig. 7. This new device, presenting a weight slightly lower than 225g in total, results a totally portable exoskeleton.

Although the redesigned prototype could not be printed out as explained in Section II, some tests of the EMG trigger and the closed-loop angular control were performed using a previous version of the exoskeleton. A picture sequence is shown in Fig. 8. Those tests point out satisfying results concerning the operation of the control architecture. In particular, Fig. 9 reports the results of the LDA classifier during one of the test. The same considerations can be extended to all studied subjects (four members of the team in total).

As it is shown, the linear boundary allows for a separation of the input data in two regions which can be associated to a particular gesture (opening and closing movements in this case). The exploited method was able to classify the correct gesture with an average accuracy of about 85%. A relevant part of miss-classifications occurred in the initial phase of both movements (highlighted with a yellow square in Fig. 9), when the very first muscular activation causes an important overlapping. Lower bounds both for sensor 1 and 2 outputs have been implemented in order to identify a "safety region" in which the command to the actuators is not associated to the classifier output. This solution led to a reduction of actuated miss-classifications but introduced a delay in actuating.

The paper presented the whole system designed to provide a hand exoskeleton to users suffering from hand reduced mobility. This new prototype is able to actuate opening and closing movements of the hand exploiting a new EMG-triggered control system. The opening gesture are driven by means of a cable transmission while the closing one is provided passively through a spring placed in a suitable position within the mechanism.

Thanks to the automatic scaling procedure and to the decision of placing only two independent servomotors on the back of the hand has led to reach a trade off between effectiveness and lightness. The proposed solution ensures a further reduced weight of the overall system exploiting topological optimization technique.

The results reported above in the paper serve as preliminary tests. At the time of writing, the new device, depicted in Section IV, has been manufactured and is involved in the testing phase at the Don Carlo Gnocchi Foundation Rehabilitation Center, Florence, Italy , of both the structural components and the control architecture.

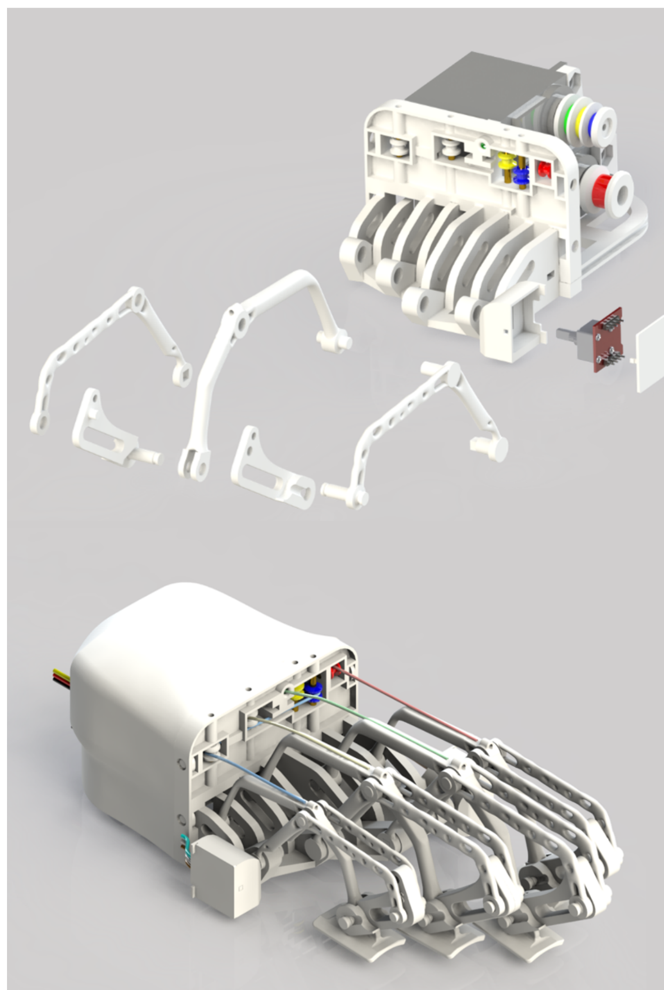


Fig. 7. Realistic render of the final version of the device

[1] *http://www.01mechatronics.com/*

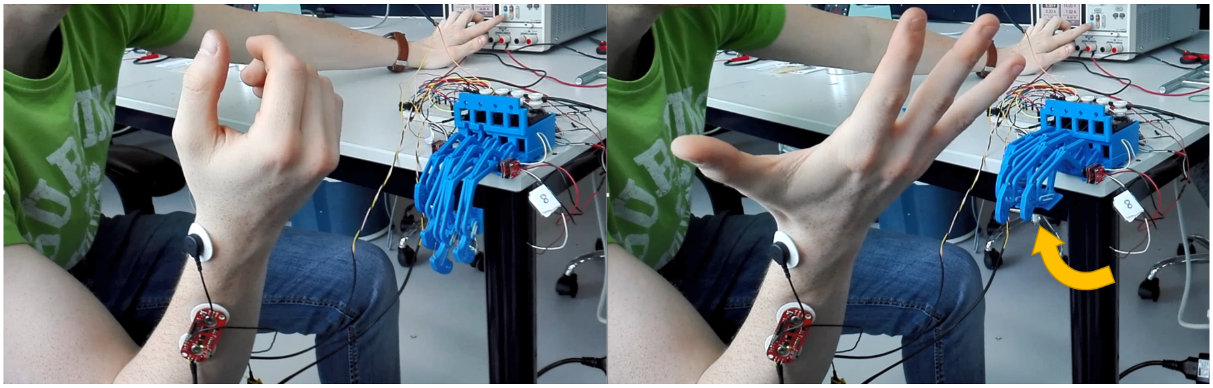


Fig. 8. Testing sequence

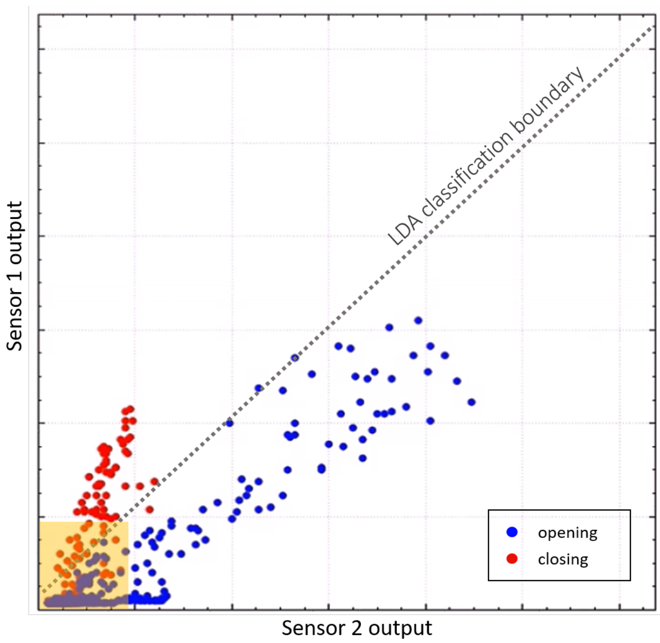


Fig. 9. Linear classification results

Acknowledgment

A special thank goes to the Don Carlo Gnocchi Foundation (Italy) which has supported this work and to the eNTERFACE'16 staff which has organized the eNTERFACE'16 Workshop.

References

[1] "Statistics for Parkinson's Disease" (2015), Parkinson's Disease Foundation. [Online]. Available: *http://www.pdf.org/*

[2] Idrogenet srl. (2014), Gloreha(R). [Online]. Available: *http://www.gloreha.com/*

[3] H. Nakayama, H. Jørgensen, H. Raaschou, and T. Olsen, "Recovery of upper extremity function in stroke patients: the copenhagen stroke study", *Archives of Physical Medicine and Rehabilitation*, vol. 75, no. 4, pp. 394-398, Apr 1994.

[4] K. Toya, T. Miyagawa, Y. Kubota, "Power-Assist Glove Operated by Predicting the Grasping Mode", *Journal of System Design and Dynamics*, vol. 5, no. 1, 2011, pp. 94-108.

[5] L. Lucas, M. DiCicco, Y. Matsuoka, "An EMGcontrolled hand exoskeleton for natural pinching", *Journal of Robotics and Mechatronics*, vol. 16, no. 5, 2004, pp. 482-488.

[6] P. Heo, G.M. Gu and S.J. Lee, "Current hand exoskeleton technologies for rehabilitation and assistive engineering", *International Journal of Precision Engineering and Manufacturing*, vol. 13, no. 5, pp. 807-824, May 2012.

[7] R. Conti, E. Meli, and A. Ridolfi, "A novel kinematic architecture for portable hand exoskeletons", *Mechatronics*, vol. 35, 2016, pp. 192-207.

[8] D. Ribas, et al., "I-AUV Mechatronics integration for the TRIDENT FP7 Project", *IEEE/ASME Transaction on Mechatronics*, vol. 20, no. 5, 2015, pp. 2583-2592.

[9] M. Fontana, M. Bergamasco and S. Marcheschi, "Haptic Hand Exoskeleton for Precision Grasp Simulation", *ASME Journal of Mechanisms and Robotics*, vol. 5, no. 4, 2013, pp. 9-18.

[10] F. Amirabdollahian et al., "Design, development and deployment of a hand/wrist exoskeleton for home-based rehabilitation after stroke - SCRIPT project", *Robotica*, vol. 32, Special Issue 8, pp. 1331-1346.

[11] A.J. Izenman, "Linear discriminant analysis", *Modern Multivariate Statistical Techniques, pp. 237280. Springer (2013).*

[12] M.P. Bendsoe, O.Sigmund, "Topology Optimization: Theory, Methods and Applications", Springer-Verlag Berlin Heidelberg, Berlin, Germany, 2003

[13] BTS Bioengineering S.p.A., Milano, Italy: http://www.btsbioengineering.com/

[14] Don Carlo Gnocchi Foundation: www.dongnocchi.it

[15] M. Luersen, R.L. Riche, and F. Guyon, "A constrained, globalized, and bounded Nelder-Mead method for engineering optimization", Structural and Multidisciplinary Optimization, vol. 27, 2004, pp. 43-54.

[16] M. Bianchi, F. Fanelli, L. Giordani, A. Ridolfi, F. Vannetti, B. Allotta, "An automatic scaling procedure for a wearable and portable hand exoskeleton", in Proceedings of 2016 IEEE International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), Bologna, Italy, Sept. 2016.

CARAMILLA - Speech Mediated Language Learning Modules

Emer Gilmartin, Jaebok Kim, Alpha Diallo, Yong Zhao, Neasa Ni Chiarain, Benjamin R. Cowan, Ketong Su, Yuyun Huang and Nick Campbell

***Abstract*—A natural model of spoken production is vital for successful acquisition of several aspects of competence in a new language. Learners need access to native speaker spoken input and written text on subjects of relevance or interest to them. We report on the motivation, design and implementation of two modules for a computer-assisted language learning (CALL) application, CARAMILLA, which is a spoken dialogue system. The two modules – a text and speech based language game to ground lexical and syntactic knowledge and a pronunciation application tailored to particular mother tongues – were created at the Enterface16 workshop held at the University of Twente, and are currently being trialled with user groups in Ireland.**

I. Introduction

Language learning is an increasingly important area of human and commercial endeavour, and has been an early adopter of various technologies, with video and audio courses available since the early days of audiovisual technology. Increasing globalisation and migration coupled with the explosion in personal technology ownership have increased the need for well designed, pedagogically oriented CALL applications. In this paper we describe the design and implementation of two CALL modules based on speech and language technology, a language game and a pronunciation practice activity, at Enterface 16, held in Twente. The modules were designed to expand a pre-existing spoken dialogue language learning agent, MILLA, developed at a previous Enterface workshop in 2014. We review the motivation for this application, describe the learner groups initially targeted, and detail the design and implementation of the two modules.

II. Motivation

Fundamental challenges to modern language learners are the development of spoken interaction skills, and the integration of lexical and syntactic knowledge to produce accurate and appropriate language in different contexts. Conversation classes and practice with native speakers are the traditional methods for development of spoken language proficiency, but classes are expensive and learners may not have access to native speakers willing to practice with them. Many current CALL (Computer Assisted Language Learning) applications address the receptive skills (listening and reading) or comprehension, much like a traditional listening or reading exercise, where the learner encounters a 'worksheet on the screen'. Pronunciation tutoring applications range from 'listen and repeat' exercises without feedback or with auto-feedback (the learner hears a recording of their attempt) to more sophisticated systems where the learner's utterance is compared with the target and feedback is given on errors and strategies to correct those errors. Examples of spoken production training based on speech technology where phoneme recognition is used to provide corrective feedback on learner input include CMU's Fluency [1], and Cabral et al's MySpeech [2]. There has been longstanding interest in the education field in general, and in language learning circles in particular, in dialogic speech activities which allow learners to engage in a coaching or testing session with a virtual tutor. Dialog systems using text and later speech have long been successfully used to tutor learners through a natural language interface in science and mathematics [3], [4]. In language learning, early systems such as VILTS [5] presented tasks and activities based on different themes which were chosen by the user, while other systems concentrated on pronunciation training via a conversational interface. Chatbot systems have been developed based on pattern matching which practice spoken interaction (or texted chat) with a conversational partner, the most difficult competence for a learner to acquire independently (e.g. Pandorabots) [6]. Recently there has been growing interest in more holistic systems where learners engage in spoken dialogue in a more naturalistic setting to roleplay tasks such as service encounters or job interviews, particularly for assessment purposes [7]. While these systems will prove vital in the acquisition and measurement of communicative competence, there is also a perpetual need for activities which combine speech technology with tried and tested language learning activities which help learners work on specific vocabulary and syntax areas. The use of gamification in educational software is currently receiving a lot of attention as a method of increasing learner motivation, and lends itself very well to language learning activities as 'grammar games' and 'information gap' activities have long been successful components of the language classroom.

The CARAMILLA project at Enterface 16, held at the University of Twente's Design Lab, was a follow on to an earlier project (MILLA) at Enterface 14, held in Bilbao. The goals of the project were to produce language learning modules to extend the capabilities of the MILLA language learning agent system. MILLA (Multi-modal Interactive Language Learning Agent) is a dialogue system providing spoken social chat at different levels (two speech-enabled web-based Pandora chatbots), pronunication and traditional grammar training [8]. MILLA was created in Bilbao by a team of nine. At the 2016 workshop, an onsite team of five postgrad students and academics in Twente supported by offsite partners in

Dublin designed and implemented an engaging and adaptable language game module (Dictogloss) and redesigned and implemented an improved pronunciation module for CARAMILLA. These activities and the original MILLA are currently being integrated with a more robust dialog platform (CARA) to form CARAMILLA. The modules are being tested by two user groups with differing needs – school language learners and teachers of Irish, and adult refugee learners of English living in Ireland.

Below we describe the learner groups targeted, and then outline the implementation of both spoken and read versions of the dictogloss activity for two languages, Irish and English, and of the pronunciation module.

III. Target Learner Groups

For our prototype system, we target two learner groups familiar to the first and fifth authors, adult refugee learners of English living in Ireland and school age learners of Irish.

A. Adult Refugee Learners of English

Learning a language to live in a country where the language is spoken is not a simple matter of attaining an academic understanding of the language– competence in spoken and written language is needed to successfully live in the new country. A major challenge in language provision to refugees is the very diverse range of backgrounds in the group. In Ireland's national language training organisation for refugees in 2008, 93 nationalities were present and learners' English language competence was very varied, with individual learners very often exhibiting a 'spiky profile' where some skills were far higher than others depending on how the learner had acquired their English [9]. This variety makes classes difficult to provide successfully as groups are extremely heterogenous. Learners would benefit greatly from individualised training as an adjunct to classes. In addition the texts underpinning language learning should address the practical needs of the learners. Better results can be expected when the texts used are tailored to practical communicative needs or areas of interest - a parent would benefit far more from an exercise on Present Tense structures based on the local education system than on a description of life on the Space Station. It is clear that flexible, free or low cost language learning resources which can be accessed from home at any time would be particularly helpful to this group. Activities which allow topics of interest to the learner to be selected as underpinning text can also foster learner autonomy and motivation [10]. Caramilla's pronunciation and Dictogloss activities are designed to cater to the individual, with English pronunciation practice available suited to learners with a variety of first languages, while the Dictogloss can be adapted to any text of interest or level of competence. The system is currently being trialled with learners working toward the CEFRL B1 level, defined as a level where the language learner 'can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.' (Council of Europe, 2001). This level was chosen as it is commonly defined as the 'threshold' for integration and relevant to all newcomers. Of course the modules can also be tailored to learners at different levels with different needs, such as academic or professional language for entry to education or work in the host country.

B. Irish language learners

Irish is a Celtic language, spoken as a community language in the Gaeltacht regions of the West of Ireland. It is an endangered minority language with very few remaining native speakers using it daily. However, Irish is an official language in Ireland and the European Union, and is taught as a required subject in all primary and secondary schools in Ireland (equivalent to K-12). Many adults worldwide learn Irish out of interest. As a minority and endangered language, there is a lack of native speaker models, and also of texts which could be used for learning purposes. The Phonetics Laboratory at Trinity College Dublin have developed a number of Irish voices for the three main dialects of the language which are available online, and are being used in CALL activities [11] including the CARAMILLA modules. In the CARAMILLA project, the spoken Dictogloss game was designed for Irish as well as for English, for use online by learners of high school age studying Irish at school, and also for teachers in training to teach other subjects through Irish in increasingly popular gaelscoileanna - schools where all subjects are taught through Irish. The availability of an application which could create spoken dictogloss exercises from Irish language texts can provide native speaker models and examples of vocabulary and syntax of interest to such learners.

IV. Dictigloss

The dictogloss is a well-known language game, which is ideal for implementation in a CALL system as it can be played by one learner with a tutor (the system) or between a group of learners with a tutor. This section describes the design and implementation of the spoken dictogloss module for the CARAMILLA system.

A. Description of Dictigloss game

A dictogloss is a complete cloze text reconstruction exercise. It is widely recognised as a useful language learning game and the version implemented here is close to that described in Rivolucri's seminal 'Grammar Games' [12]. In the game learners read or listen to a text once, and then reconstruct some or all of it - by filling in blanks on a worksheet or by writing and rewriting fragments to assemble a coherent text. Dictogloss grounds syntax and vocabulary knowledge 'by stealth', and as a listening exercise, also aids in developing listening skills. A major advantage of the exercise is its flexibility - different areas of vocabulary and syntax can be addressed at different levels of difficulty by simply changing the text used in the exercise. Thus, the game can be very useful for learners who need language for practical purposes - for example, nurses hoping to work in an English speaking country can improve necessary vocabulary using a description of clinical practices, while refugees learning the past tenses could reconstruct an account of a well-known historical incident in the host country drawn from a history resource or a current news story of interest scraped from a news website.

The typical procedure for a spoken dictogloss exercise is as follows:

1) Teacher asks learners to relax, clear desks, and listen - there is no note-taking allowed
2) Learners listen as teacher reads text at normal speed with normal intonation
3) When text is finished, learners, alone or in teams write down words remembered from text
4) Teacher distributes a blank rectangular grid containing enough cells for all the words (tokens) in the text.
5) Learners take turns guessing words that were in text, teacher tells them grid co-ordinates and learners enter correct guesses into grid. Points are awarded for correct words.
6) Teacher provides support in the form of hints and encouragement as learners complete grid

After they write down the words that they remember in their correct places, learners start to reconstruct the sentences in the text by guessing words which could appear in the text. Seeing the string 'I __ a __ car', they will realise that a verb is needed in the first blank and will start guessing verbs appropriate to the world of cars. The second blank will elicit adjectives. This hypothesis and test stage is the core of the exercise allowing learners to consolidate their syntactic and lexical knowledge in an interactive context. The spoken dictogloss was implemented in JAVA as part of CARAMILLA as described below.

B. JAVA Implementation of Dictogloss

The dictigloss process was implemented in Java, by creating web pages using Java Servlets and JSP, as a game in the CARAMILLA computer assisted language learning (CALL) system. Both English and Irish versions were implemented, with the only differences between the two being the source where the text was retrieved from, in this case the English Simple Wikipedia and Irish Wikipedia pages, and the text-to-speech (TTS) programs used to read the text to the user.

The advantage of Simple Wikipedia is that the text contains short sentences using a basic lexicon, so that it can more easily be understood by learners. Another advantage of these shorter sentences and simple vocabulary, unlike the heavily embedded phrases found in Wikipedia itself, is that they are closer to speech in syntax and lexicon and are thus more suitable for spoken delivery.

The requirements for the text-to-speech (TTS) systems were that they produce high quality speech in an accent relevant to the user groups. For English the system used was CereProc's Caitlin voice for the English version [13], which is a pleasant highly intelligible female voice with an Irish accent. For Irish, we used Irish voices from the Abair synthesizer created by the Phonetics and Speech Laboratory, Trinity College Dublin [14], which include male and female voices from the three main dialects of the language. In our system, for both English and Irish, we synthesise the chosen text by sending TTS queries to web servers in Trinity College Dublin which return the synthesis results. Since we have very long texts for the Dictogloss game module, it can be very time consuming to send the complete TTS query to the web server, and a lag is introduced at run time. In addition, if we send the whole text content for synthesis, there will be no significant pause between each sentence, which sounds unnatural and makes comprehension difficult. Therefore, we split the long text into single sentences, which are sent for synthesis sequentially. This method has the added advantage of introducing a short pause between sentences, more closely modelling natural speech.

In this implementation of dictigloss, the game is browser-based and is accessed from CARAMILLA's main menu. On the menu page, the user selects Dictogloss and chooses whether they want to listen to or read the text and then choose a topic from a currently predefined list. In the future the list could just be a list of subjects that may be of interest to the user and when they choose the subject current articles could be scraped from news websites or journals and returned. Once the user has chosen an action (read or listen) and a predefined topic, the text is scraped from the English or Irish Wikipedia page and a predefined number of sentences are formed for the user. If the user chooses to read the text, text will be displayed for them, but if they chose to listen, the TTS system will be called and will synthesize the text line by line as described above and read it out at normal speed.

Once the user has finished reading or listening, they are taken to a page where they are presented with an empty grid representing the text. The grid comprises underscores indicating each word and is prepopulated with the punctuation marks from the original text, thus indicating where sentences end. This greatly aids learners and keeps the difficulty level of the game to an intermediate level. Of, course this step could be made optional in later versions to provide a more challenging exercise. The interface contains a textbox where the user can enter words that they think are in the text. The user is invited by the system to guess a word from the text, and then types the word into the box. If the word that the user types, regardless of case, is in the text, the word replaces the underscores in all the places it appears in the original text. When the user guesses the correct word they get points for the word. Currently, each word is only worth a single point, but in the future words could have different points values based on difficulty - for example function words such as 'a' or 'then' appear in most texts and could carry lower marks. While the user is going through the process of inferring which words come next, they may get stuck or run out of ideas. If this happens, the user can ask for help just as they would if they were going through the process with a teacher. By entering the text 'hint' the user is currently given the missing word, but this is being extended to allow the system to give the user the definition of the word or a synonym of the word by consulting an online dictionary or thesaurus. In the real life game, learners say their guesses but this is challenging to implement using ASR as single words do not provide much data for accurate recognition. Also, learners would be apt to produce guesses embedded in spoken phrases which would require extra analsysis work by the recogniser. A spoken guess option will be explored in later versions.

V. Implementation of Pronunciation Module

The pronunciation training module is based on the use of automatic speech recognition to compare a learner's production of an utterance with a model. The intention is to provide pronunciation training for learners of different first languages (L1), as learners from different L1s speakers are known to make characteristic errors.

A. Corpus of typical errors for pronunciation module

Learner of a second language often produce pronunciation errors due to interference from the phoneset of their first language or the presence of phonemes in the target language which do not exist in their mother tongue. For example Spanish speaking learners of English often have difficulty with the short vowel in 'fill', pronouncing it as a longer i: sound as in 'feel'. Errors characteristic to English learners from different first languages are well known to teachers and knowledge of such errors can usefully tailor pronunciation training to a learner's needs. To personalise the pronunciation module, we decided to include this knowledge in the design of pronunciation example. Thus, to feed our pronunciation module with some data, we created a corpus of 50 practice sentences with common pronunciation mistakes in English by language background. Pairs of frequently confused English language phonemes for learners from 10 first languages (Arabic, Chinese, Croatian, Dutch, Finnish, French, German, Korean, Spanish, and Turkish) were collected from an website maintained by an expert teacher of English pronunciation [15]. A carrier sentence for each phoneme pair was then created for use in the application. For each sentence, we created a tip for how to pronounce the commonly mistaken phoneme pair and a phoneme transcription with lexical stress from the CMU pronouncing dictionary [16]. For example, for Spanish or French speakers, the phrase 'These shoes fit my feet' containing the **I vs i:** sounds in the minimal pair 'fit' and 'feet' was used to test these commonly confused vowel sounds. The pronunciation system interface was designed to show the carrier sentence, the pronunciation tip, and a line drawing of the normal positioning of articulators in the vocal tract during native pronunciation of the sound in question.

VI. Creation of Pronunciation Scoring Application

Pronunciation scoring is generally utilised in language learning applications to obtain global scores which tells overall goodness of proficiency on an utterance. However, global scores do not give specific information of where students make mistakes, which results in less useful applications [17]. Hence, pronunciation testing should give not only global scores but also local scores at phoneme levels, and learner should pay attention to which phonemes they cannot pronounce correctly.

One of widely used methods to detect phoneme level errors is Confidence-Measure (CM) based error detection, which is often integrated with Hidden Markov Model (HMM) [18]. A practical advantage of this approach is to utilise easily Automatic Speech Recognition (ASR) system. Using speaker adaptation techniques such as Maximum Likelihood Linear Regression (MLLR). However, the CM-based approach uses the same feature set for all the phones, which might not be an optimal approach to detect explicit errors on a specific phoneme. Moreover, Speaker adaptation can bring over-fitting to a target user, which makes confidence scores even more unreliable.

In contrast, Linear Discriminant Analysis (LDA)-based classifiers optimise acoustic feature sets for each phoneme, which results in significant improvement of error detection on particular phones [17]. However, their evaluations are quite limited to small Dutch phoneme sets, which cannot be generalised.

Our approach is based on Witt's goodness of pronunciation (GoP) [18]. Figure 1 describes the basic algorithm. This algorithm calculates a distance between answer (log likelihood of forced-alignment results) and target (that of phoneme recogniser) as follows:

$$GOP_1(q_i) = \frac{\log\left(p(O|q_i)\right)}{NF(O)} - \frac{\max_{j=1}^{J}\log\left(p(O|q_j)\right)}{NF(O)}| \quad (1)$$

where q_i is i'th phoneme in an utterance and $NF(O)$ is the sum of log-likelihood of all frames in the observation.

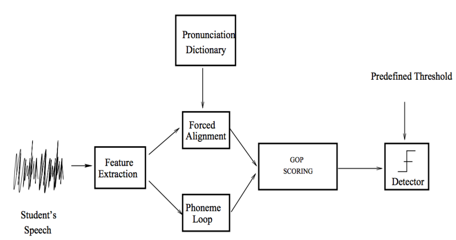


Fig. 1. *Basic algorithm of GoP [18]*

Again, this approach can measure local score for each phoneme but it does not indicate explicit errors (e.g. b → v). Hence, we modelled explicit error networks for mother languages of users (e.g. Chinese, Korea, Arabic, and etc.). For this, we collected utterances where non-native speakers make common mistakes in general from a user group during the eNTERFACE workshop. Figure 2 depicts an example of the error networks.

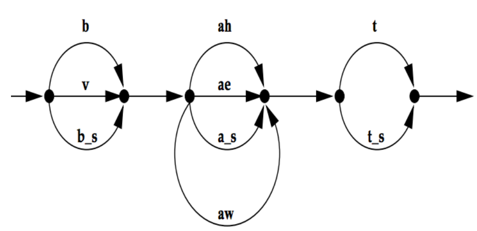


Fig. 2. *An example of error model [18]*

The following equation is to combine a local score and a score of its explicit error model:

$$GOP_2(q_i) = GOP_1(q_i) + KGOP_e(q_i) \quad (2)$$

where K is a scaling factor. The details to calculate $GOP_e(Q_i)$ can be found in [18].

In addition, our system explicitly displays which kinds of errors the user makes, which are categorised into deleted (missing), substitute, inserted as the same way as used in the evaluation of ASR system. Hence, users can realise both their proficiency and explicit errors.

For reproduction, we implemented our system using HTK [19] and Sphinx 4 [20] tool kits. For the robust recognition in the wild, we employed sub-band OSF-based voice activity detection [21].

VII. Conclusion and Future Work

The four-week workshop resulted in the implementation of prototype modules for Dictogloss and pronunciation. Subsequent to the workshop, we have started a testing and evaluation process with real users in the target groups, and are working on improvements to the sytem. The Dictogloss module is currently being extended to incorporate texts of different levels and on different subjects in both English and Irish, and is being trialled with two groups of real users. The English language version is being piloted with refugee learners in a centre for language and integration courses in Dublin, while the Irish language version will be piloted with Irish language students in Trinity College Dublin in the coming academic year. A spoken input version will be explored as will several minor improvements giving options for spellchecking, thesaurus access, and switching between punctuation provided and completely blank grids to vary the challenge. For pronunciation tutoring, apart from getting a score from GOP and seeing the static line drawing and tip on the screen it could be very beneficial for the learner to learn how to pronounce by seeing the lip movement from a virtual agent. In addition, when students are playing the Dictogloss game, having a virtual agent to read the story and even to talk with may increase engagement and indeed retention of lexical and syntactic items encountered during the sessions. We are therefore exploring the open source animation toolkit Smartbody [5] to build a virtual avatar for our CARAMILLA system. Lip-synchronizing and facial expression can be handled given a synthesized audio file. We will integrate this into the Java web-based framework in the future. We also plan to expand the range of texts available to the user in both languages, and ultimately hope to automatically source texts based on keywords entered by any user. Our longer term goal is to incorporate a virtual learner portfolio, tied to the Common European Framework, which would allow learners to track their progress and plan their future activity, thus fostering autonomous learning and also mapping language acquisition to an internationally recognised scale.

Acknowledgment

The authors would like to thank the organizers of Enterface 16 and the University of Twente for making this work possible. This work is supported by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences and Technologies ERA-NET (CHISTERA) JOKER project, JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot, and by the Speech Communication Lab, Trinity College Dublin.

References

[1] M. Eskenazi and S. Hansma, "The fluency pronunciation trainer," in *Proceedings of the STiLL Workshop*, 1998.
[2] J. P. Cabral, M. Kane, Z. Ahmed, M. Abou-Zleikha, E. Székely, A. Zahra, K. U. Ogbureke, P. Cahill, J. Carson-Berndsen, and S. Schlögl, "Rapidly testing the interaction model of a pronunciation training system via wizard-of-oz." in *International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 4136–4142.
[3] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue," *Education, IEEE Transactions on*, vol. 48, no. 4, pp. 612–618, 2005.
[4] D. J. Litman and S. Silliman, "ITSPOKE: An intelligent tutoring spoken dialogue system," in *Demonstration Papers at HLT-NAACL 2004*, 2004, pp. 5–8. [Online]. Available: http://dl.acm.org/citation.cfm?id=1614027
[5] M. E. Rypa and P. Price, "VILTS: A tale of two technologies," *Calico Journal*, vol. 16, no. 3, pp. 385–404, 1999.
[6] R. S. Wallace, *Be Your Own Botmaster: The Step By Step Guide to Creating, Hosting and Selling Your Own AI Chat Bot On Pandorabots*. ALICE AI foundations, Incorporated, 2003.
[7] K. Evanini, S. Singh, A. Loukina, X. Wang, and C. M. Lee, "Content-based automated assessment of non-native spoken language proficiency in a simulated conversation."
[8] J. P. Cabral, N. Campbell, S. Ganesh, E. Gilmartin, F. Haider, E. Kenny, M. Kheirkhah, A. Murphy, N. N Chiarin, T. Pellegrini, and others, "MILLA Multimodal Interactive Language Learning Agent," in *DialWatt - Semdial 2014*, 2014.
[9] E. Gilmartin, "Language training for adult refugees: The integrate ireland experience." *Adult Learner: The Irish Journal of Adult and Community Education*, vol. 97, p. 110, 2008.
[10] D. Little, "Language learner autonomy: Some fundamental considerations revisited," *International Journal of Innovation in Language Learning and Teaching*, vol. 1, no. 1, pp. 14–29, 2007.
[11] A. N. Chasaide, N. N. Chiaráin, C. Wendler, H. Berthelsen, A. C. Kelly, E. Gilmartin, E. N. Dhonnchadha, and C. Gobl, "Towards personalised, synthesis-based content in irish (gaelic) language education." in *SLaTE*, 2011, pp. 29–32.
[12] M. Rinvolucri, "Grammar games: Cognitive," *Affective, and Drama Activatiobn for EFL Students*, 1984.
[13] "CereVoice Engine Text-to-Speech SDK | CereProc Text-to-Speech," 2014. [Online]. Available: https://www.cereproc.com/en/products/sdk
[14] "abair.ie The Irish Language Synthesiser." [Online]. Available: http://www.abair.tcd.ie/
[15] "Teaching and learning English pronunciation." [Online]. Available: http://www.tedpower.co.uk/phono.html
[16] "The CMU Pronouncing Dictionary." [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[17] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.
[18] S. M. Witt, *Use of speech recognition in computer-assisted language learning*. University of Cambridge, 1999.
[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.
[20] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," 2004.
[21] J. Ramírez, J. C. Segura, C. Benítez, A. De la Torre, and A. Rubio, "An effective subband osf-based vad with noise reduction for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1119–1129, 2005.

Index of Authors