A demonstration of audiovisual sensitive artificial listeners

Marc Schröder, Elisabetta Bevacqua, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Björn Schuller, Etienne De Sevin, Michel Valstar, Martin Wöllmer

Angaben zur Veröffentlichung / Publication details:

Schröder, Marc, Elisabetta Bevacqua, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Sathish Pammi, et al. 2009. "A demonstration of audiovisual sensitive artificial listeners." In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 10-12 September 2009, Amsterdam, Netherlands*, edited by Jeffrey Cohn, Anton Nijholt, Maja Pantic, Ferdinand Beljaars, and Pyrrhos Stathis. Piscataway, NJ: IEEE. https://doi.org/10.1109/ACII.2009.5349530.

A Demonstration of Audiovisual Sensitive Artificial Listeners

Marc Schröder*, Elisabetta Bevacqua°, Florian Eyben+, Hatice Gunes□, Dirk Heylen#, Mark ter Maat#, Sathish Pammi*, Maja Pantic□, Catherine Pelachaud°, Björn Schuller+, Etienne de Sevin°, Michel Valstar□, Martin Wöllmer+

*DFKI GmbH, Saarbrücken, Germany

°CNRS, Telecom ParisTech, Paris, France

+Technische Universität München, Munich, Germany

□Imperial College London, London, United Kingdom

#Universiteit Twente, Twente, The Netherlands

schroed@dfki.de

Abstract

Sensitive Artificial Listeners (SAL) are virtual dialogue partners who, despite their very limited verbal understanding, intend to engage the user in a conversation by paying attention to the user's emotions and non-verbal expressions. The SAL characters have their own emotionally defined personality, and attempt to drag the user towards their dominant emotion, through a combination of verbal and non-verbal expression. The demonstrator shows an early version of the fully autonomous SAL system based on audiovisual analysis and synthesis.

1. Introduction

The Sensitive Artificial Listener demo is an early version of a system supporting sustained emotionally-coloured machine-human interaction using non-verbal expression, which is currently being developed by the FP7 project SEMAINE. The system aims to engage the user in a dialogue by paying attention to the user's non-verbal expression, and reacting accordingly.

The motivation for the system comes from the observation that existing multimodal dialogue systems usually lack the "soft skills" that humans naturally use to indicate to each other that they are interested in a conversation with each other, that they are listening, that they want the speaker to keep on talking or that they want to start speaking themselves. What humans do without any specific effort however is too difficult for today's interactive technology.

To simplify the challenge somewhat, the SAL system avoids task-oriented dialogue. Instead, it models the type of interaction found at parties: you listen to someone you want to chat with, and without really understanding much of what they are saying, you exhibit all the signs that are needed for them to continue talking to you. Similarly to the Rapport agent [1], SAL characters show non-verbal listener signals; in addition, they can also speak to engage the user in a simple dialogue.

The approach has been test-run using Wizard of Oz setups at various stages of maturity [2][3]. This has allowed us to fine-tune the scripts used by the various characters, in order to react to the emotional state of the user in plausible ways despite the lack of verbal knowledge.

The system is being developed in large parts as open source. A first public version of the system has been made publicly available on sourceforge [4][5]. It provides the SEMAINE API, a re-usable middleware for emotion-oriented systems, as well as elementary system components. The next public version, which will correspond to an enhanced version of the current demo, is scheduled for publication by the end of 2009.

The system is based around standard representation languages for the communication among its modules, thereby promoting interoperability and reuse.

2. Demonstration setup

The setup is a simple face-to-face setup, with one human user sitting in front of a computer screen showing the face of an Embodied Conversational Agent (ECA). The user is wearing a headset for voice analysis and is recorded by a video camera for facial expression analysis. The ECA is speaking through loudspeakers, and is showing both verbal and non-verbal behaviour.

A second computer screen shows a system monitor, displaying graphically the current information flow in the system, and providing detailed information about the processing times used by components and the data flowing between components.

3. Technical content of the demonstration

Technically, the demonstrator system is a multimodal interactive system with components integrated across programming languages and operating systems by means of a middleware layer, the SEMAINE API [4]. Components may be running on Windows, Mac OS X or Linux, and are programmed in either C++ or Java.

User behaviour is first analysed in terms of low-level *feature extractors* for speech and face (Figure 1). The resulting feature vectors are processed by *analyser* components, in terms of context-free behavioural and mean-

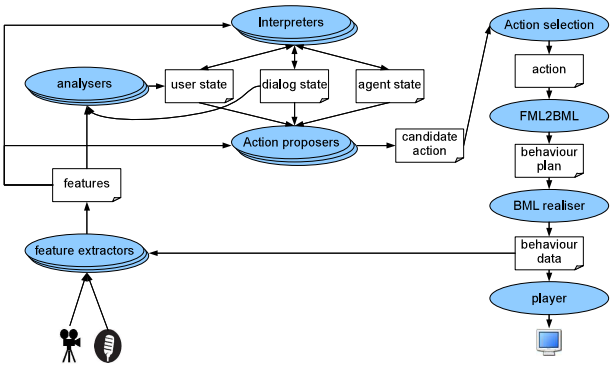

Figure 1: Conceptual architecture of the SAL system.

ing analyses. These include basic aspects such as the question whether the user is currently speaking or silent, behavioural interpretations such as whether the user is nodding or shaking the head, and higher-level analyses such as the user's arousal, valence and interest. Furthermore, keyword spotting technology is used for capturing at least some aspects of the verbal content.

A range of *interpreter* components is used to make context-dependent judgements based on the evidence from the analysers. Here, collected evidence for a certain emotional state becomes the system's "current best guess" of the actual state of the user and the dialogue. Furthermore, interpreters deduce an agent state, such as the agent's interest in talking more with the user given the dialogue history and the current user state.

Drawing on that collected evidence, a set of *action proposers* produces candidate actions, which can be non-verbal actions such as a smile mimicking the user's smile, backchannels while the user is speaking, or verbal contributions.

An *action selection* component attempts to prioritise actions where several candidates are produced at the same time.

Some of the actions being represented in terms of their functions, they are mapped onto suitable behaviour by a *function-to-behaviour* component using a multimodal mapping table, listing the known behaviours in different modalities that can be used to realise a given function.

Behavioural representations are then processed by *speech synthesis* and a *visual behaviour realiser*, and rendered by a *player*.

Details on the technological setup are available in [4].

4. Outline of a SAL session

After initial explanations, the user chooses an interaction partner among, Spike, Poppy, Obadiah and Prudence (see Figure 2).

Each character has its own emotionally coloured scripts, covering among themselves the four quadrants of activation-evaluation space. Poppy is positive-active, i.e. a cheerful character that will insist on seeing the bright side of things. Spike is negative-active, showing aggression unless the user is also in an aggressive mood, in which case he would consent that that is the right attitude. Obadiah is negative-passive, or gloomy, and will basically suggest to the user that there is no way out of the difficulties that the user undoubtedly is in. Finally, Prudence is pragmatic, or positive-passive, and will attempt to positively calm down the user in a way that is oriented towards solutions.

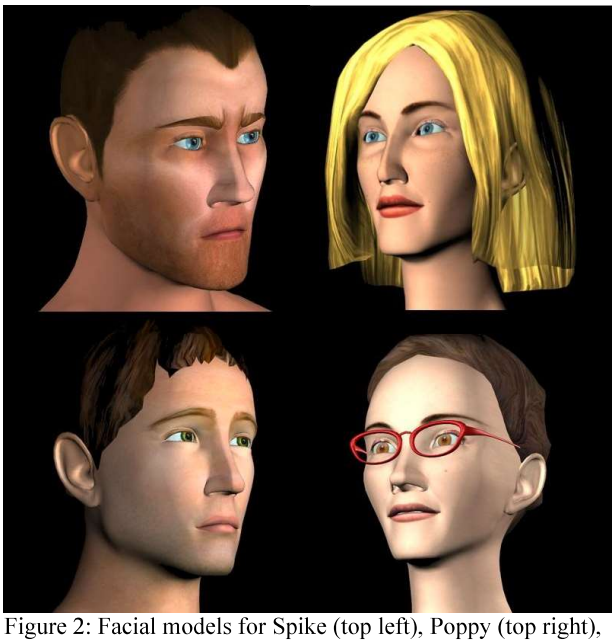
Figure 2: Facial models for Spike (top left), Poppy (top right), Obadiah (bottom left) and Prudence (bottom right).

Over the course of a session, the user will first choose to talk to one of the characters. That character will use its script to try and sustain the conversation, until either the user or the agent decides that the user should talk to another character. In the lab, sessions typically last for around 20 minutes; in the demo session, much shorter sessions with changing users are anticipated.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE).

References

[1] J. Gratch et al., "Creating Rapport with Virtual Agents," *Intelligent Virtual Agents*, 2007, pp. 125-138; http://dx.doi.org/10.1007/978-3-540-74997-4_12.

[2] E. Douglas-Cowie et al., "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," *Affective Computing and Intelligent Interaction*, 2007, pp. 488-500; http://dx.doi.org/10.1007/978-3-540-74889-2_43.

[3] E. Douglas-Cowie et al., "The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation," Marrakech, Morocco: 2008, pp. 1-4.

[4] M. Schröder et al., *SEMAINE deliverable D1b: First integrated system*, 2008; http://semaine.sourceforge.net/SEMAINE-1.0/D1b First integrated system.pdf.

[5] "SEMAINE sourceforge page"; http://sourceforge.net/projects/semaine/.