A Multimodal Listener Behaviour Driven by Audio Input

Etienne de Sevin[1], Elisabetta Bevacqua[1], Sathish Pammi[2], Catherine Pelachaud[1],
Marc Schröder[2], Björn Schuller[3]
[1] CNRS - LTCI Telecom ParisTech, 37/39, rue Dareau, 75014 Paris, France
firstname.lastname@telecom-paristech.fr
[2] DFKI GmbH, Language Technology Lab, Saarbrücken and Berlin, Germany
firstname.lastname@dfki.de
[3] Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany
lastname@tum.de

ABSTRACT

Our aim is to build a platform allowing a user to chat with virtual agent. The agent displays audio-visual backchannels as a response to the user's verbal and nonverbal behaviours. Our system takes as inputs the audio-visual signals of the user and outputs synchronously the audio-visual behaviours of the agent. In this paper, we describe the SEMAINE architecture and the data flow that goes from inputs (audio and video) to outputs (voice synthesizer and virtual characters), going through analysers and interpreters. We focus, more particularly, on the multimodal behaviour of the listener model driven by audio input.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents.

General Terms

Algorithms, Experimentation, Human Factors, Standardization.

Keywords

Description: Application, Inspiration: Artificial Intelligence, Focus Keyword: (Virtual) Agents (any subarea), Real-Time, ECAs.

1. INTRODUCTION

Our work is part of the EU project SEMAINE[1], which aims at building an agent called SAL (Sensitive Artificial Listener), endowed with social capabilities that it applies to sustain an emotionally coloured conversation with a human user. SAL characters are used to collect data about human interaction, they incite users into talking by simply appearing to listen and encouraging now and then with pre-defined phrases. To this end a credible listening behaviour is important.

In the SEMAINE architecture, the data flow comes from inputs (audio and video), passes through analysers and interpreters, then some actions are proposed and selected to be sent to the generation of the agent animation. The information from inputs can be at the level of features (energy of speech) or synthesized at the level of the user state (pitch direction). This information is used by the output part of the system to generate the animation of the virtual agent. In this article, we focus on the generation of adapted listener behaviours based on audio inputs to enhance the interaction with the user.

2. STATE OF THE ART

Previous research has approached the issue of implementing interactive Embodied Conversational Agents (ECA) endowed with believable listener behaviour: Kopp et al. [1] proposed a model to generate the agent's behaviour taking into account how it feels and what it thinks about the speaker's speech. This model has been tested with Max, a virtual human based on a BDI architecture [2]. Max is able to display multimodal backchannels (like head nod, shake, tilt and protrusion with various repetitions and different movement quality) that are triggered solely according to the written input that the user types on a keyboard. To determine backchannel timing, the system applies an end-of-utterance detection, since listeners' signals are often emitted on phrases boundaries. So far the end of an utterance is signalled when the "enter" key is typed. Other appropriate moments to provide backchannels are found by considering the type of word pronounced by the user: a backchannel is more appropriate after a "relevant" word, like a noun or a verb, than after an article.

Gratch et al. [3] developed the "Rapport Agent" that is based on a modular architecture. This agent, able to provide non-verbal backchannels, was implemented to study the level of rapport that users feel while speaking to a virtual agent. The system analyses the user's non-verbal behaviour (nod, shake, head movement, mimicry) and some features of the user's voice to decide when a backchannel must be displayed. Backchannels comprehend signals like head nods, head shakes, head rolls and gaze shifts.

We aim to build a modular architecture that is a fully operational implementation of an autonomous conversational agent according to the SAIBA standard [4]. Our system takes into account both the content of the user's speech and his acoustic and visual behaviour to determine the multimodal backchannel timing and its final shape. Moreover different types of behaviour can be generated according to the type of agent.

3. SEMAINE ARCHITECTURE

The system architecture is implemented on the basis of the SEMAINE API, a distributed component integration middleware. We briefly describe the principles of that platform before describing the components making up the system.

3.1 Component integration middleware

Our system is implemented on top of the SEMAINE API, a distributed multi-platform component integration framework for real-time interactive systems [5].

[1] http://www.semaine-project.eu/

The communication passes via the message-oriented middleware ActiveMQ, which is reasonably fast and supports multiple operating systems and programming languages. For component integration, the SEMAINE API encapsulates the communication layer in terms of components that receive and send messages, and a system manager that verifies the overall system state and provides a centralised clock independent of the individual system clocks. The API makes it particularly easy for components to communicate via a number of standard representation formats such as the Behaviour Markup Language (BML), but also allows for arbitrary messages so that the functionality can be easily extended. The platform is publicly available as open source; detailed information about its extensibility is available [5].

3.2 Architecture of the SAL system

The conceptual architecture of our system is shown in Figure 1. Components are shown as blue ovals, message types as white rectangles. The raw user input is converted by a set of feature extractors into raw feature vectors which are sent very frequently (e.g., every 10 ms for audio, and for every video frame). We call *analysers* components such as classifiers which derive some sense from the raw features in a context free manner; *interpreters* are then considering the analysis results in the light of everything the agent knows about the current and recent state of the world, and ultimately derive the system's “current best guess” regarding the state of the user and the dialogue, and update the agent's own state in the light of this evidence. In parallel, *action proposers* can continuously generate candidate actions, which are filtered by an *action selection* before they are realised as agent behaviour. In order to realise an action, the multimodal behaviour is *planned* based on a representation of the communicative function, *realised* in terms of the synthetic audio and synchronised player directives for the visual behaviour, and finally it is rendered by a 3D character *player*.

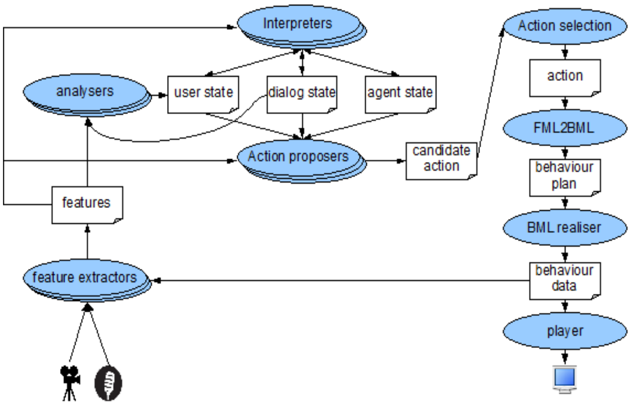

Figure 1. Simplified architecture of the SEMAINE system.

3.3 Callbacks

We have implemented callbacks in the SEMAINE project to inform other modules about the processing state of a certain piece of information. Specifically, the 3D player has to send when the agent begins or finishes a specific animation or utterance thanks to a unique ID system. This information is used in several places in the system, in particular:

- The speech analysis modules use the messages regarding start and end of audio playback to prevent voice activity detection while the agent is speaking.
- The Action Selection module avoids sending additional playback requests (backchannels or utterances) while the system is still playing an animation, which would interrupt the current system output.

4. AUDIO INPUT

Three types of information based on acoustic analysis are used for the agent control at the moment: speech pauses, pitch, and the level of the user’s interest. These will be described in short in this section. The used audio analysis components are part of the open source SEMAINE system (for an overview, refer to [6]). Audio is generally processed by a pre-emphasis with a factor of 0.97 and 25 ms frames by Hamming windows each 10 ms - except for pitch analysis where 40 ms and no pre-emphasis are used.

4.1 Speech Pauses

To measure the length of speech pauses, we use voice activity detection (VAD) that combines two algorithms to robustly detect voice activity even in noisy environments. First, an adaptive rule-based VAD analyses the low-level descriptors logarithmic energy, line spectral pair (LSP) frequencies (derived from linear prediction coefficients to determine a deviation factor), and the Mel-spectrum to calculate its entropy. The LSP deviation computation is based on the assumption that for unvoiced signals and silence the line spectral frequencies are distributed equally among the unit circle in the complex plain, i.e. distributed equally over the range [0, π]. For voiced segments one or more LSP frequencies differ from their expected position. Thus, the deviation factor is computed as the sum of squared differences between the actual and the expected positions. The raw contours of the three parameters are then smoothed using an asymmetric first order recursive exponential filter. With a 10 ms frame rate, the smoothing factor alpha is 0.05 for rising signals (i.e. the current value is greater than the last filtered value) and 0.4 for falling signals.

The actual VAD is then performed using a fuzzy logic: for each parameter five thresholds are computed as the parameter's current mean plus/minus N times the parameter's standard deviation (these statistics are initialized from the first second of input data and then continuously updated). If the current parameter value is higher than a threshold N, the fuzzy score for this parameter is set to N*0.2. For the final decision a weighted sum of the three individual fuzzy scores is computed, exponentially smoothed, and a threshold of 0.5 is applied for the final binary voicing decision. The binary output of the rule based VAD is used to train a model of the speaker's voice and the background noise. Mel-Frequency cepstral coefficients (MFCC) are used as input features to the model, which is a simple nearest neighbour model containing the mean values of the input features for each class (voice / non-voice). An unknown input is assigned the class with the smallest distance between the class’s mean vector and the input. This simple classifier allows for very fast and efficient on-line model updates and on-line classification. After the two models for background noise and speaker voice have been trained with data of at least 5 seconds each, the system stops using the rule-based voice activity input. The classifier output is now used in a feedback loop as training label for new data, and the classifier output replaces the rule-based VAD output as the final voicing decision. If the final voicing decision changes state, it remains at least two frames in the novel one.

4.2 Pitch

Pitch is determined by a combination of Auto Correlation Function (ACF) and Cepstrum-based analysis: the voicing probability estimation is based on the ACF's maximum in the range of 50-500 Hz divided by the value at the origin; the frequency is based on maximum peak picking in the Cepstrum (equal range). An unvoiced segment is assumed for the voicing probability being lower than 0.55, whereby outliers between two neighbouring frames are smoothed by forced state change.

4.3 Level of User's Interest

To determine the Level of User's Interest (LUI) we use the Audio-Visual Interest Corpus' (AVIC) [7] for acoustic model construction and feature selection. 3,002 turn instances are used as in [8] annotated by four annotators individually (906 are thus discarded due to an ambiguous inter-labelling, further 6 due to their shortness). After clustering of the five annotated ordinal classes to three, the final classes (ID, instances) are disinterest/indifference/neutral ("LUI -1", 553, i.e. low interest), interest ("LUI 0", 2,279), curiosity ("LUI 1", 170, i.e. high interest). Note that in this work we do not aim at provision of recognition results, but to obtain the best possible model for the system. Thus, the complete set as named is used for feature selection and model training. We select 127 acoustic features from 3120 obtained by systematic generation. The running system extracts the full set, as these features are partly needed elsewhere, and picks the needed from a look-up table at an overall real-time-factor < 0.02. All data is normalised based on training, and linear Kernel Support Vector Machines are used that provide a probability estimate. A continuous interest estimate in [-1,1] is obtained by computation of the center of gravity using the sum of the probability estimates multiplied with each class's index as ordinal value (cf. above). This led to better results over use of Support Vector Regression in our case, as by four labellers' average LUI the continuum as needed for a regression model is not well covered in the range. For a rough impression on performance, the reader is referred to [7] where speaker independent results on the same instances reach 68.6% (3 classes) and 77.5% (2 classes) in a cyclic five-fold leave-one-speaker-group-out evaluation.

5. AUDIO-BASED BEHAVIOUR MODEL

Audio inputs are used to manage the listening behaviour of the virtual agent at several levels of the architecture. The output callbacks are also useful for the audio input modules. Research has shown that there is a strong correlation between backchannel signals and the verbal and non-verbal behaviours performed by the speaker. From the literature [9, 10] we have built some probabilistic rules to decide when a backchannel signal should be triggered. Our system analyses speaker's behaviours looking at when a given rule is satisfied that could prompt an agent's signal; for example, a head nod or a variation in the pitch of the user's voice will trigger a backchannel with a certain probability.

5.1 Speaker vs. listener

So far SEMAINE does not model interruption. A model of turn-taking [11] determines when the agent should speak or listen according to the interaction. The audio analysis sends information when a speaker has finished the turn. It is stored in the user-state data. Then, the agent changes role from listener to speaker. The listener intent planner is used to trigger backchannels only in listener mode while the listener action selection module controls which backchannels or utterances are to be displayed by the agent.

5.2 Backchannel triggers

The backchannel trigger rules are defined through an XML-based language and are written in an external file uploaded at the beginning of the interaction. The definition of a rule is a triplet:

RULE = (name; usersignals; backchannels);

in which:

- *name* is the unique name of the rule. For example the name of the following rule is trigger-head nod:
- *usersignals* is the list of the user's signals that must be detected to trigger the rule;
- *backchannels* contains the possible listener's backchannels that can be generated with a certain **probability** when the rule is applied. The agent can either mimic the user's non-verbal behaviour (when listed in the *usersignals* tag) or emit a response/reactive backchannel. Mimicry behaviour has been considered that such a behaviour not only often occurs during human-human interaction, but it also has an important role in the successfulness of the interaction [12].

We adopted an XML-based language since it has several advantages: the set of rules can be easily modified or extended without requiring any change in the source code.

Audio input (as described in Section 4) is used in our system to drive the agent's listening behaviour. Pauses or pitch changes in the user's voice can trigger a backchannel signal, for such a reason appropriate rules have been created. For example, when a pause longer than 110 ms is detected, a backchannel is prompted with a probability of 0.95 [10].

Studies have been performed to evaluate how users perceive the listening agent's behaviour during an interaction [13], and their results have shown that users had the impression that the agent was actually listening.

5.3 Backchannel selection

The selection algorithm [14] has to choose between two types of backchannels: mimicry and response backchannels. Mimicry is chosen preferentially when the agent perceives that the user is very interested in the interaction so that the agent can show its high engagement in the interaction [3]. The response backchannels, showing the communication intentions of the agent, are used when the ECA detects that the user loses interest in the interaction. They are also used to encourage the user to be interested in the interaction [15].

The user's estimated interest level is calculated from user's speech (see section 4.3). We define the relation between backchannel priorities and the Level of User's Interest as follows: when the ECA estimates that the interest level of the user is near to the maximum (LUI > 0.75), mimicry is chosen preferentially. If the agent detects that the user begins to be less interested (LUI < 0.75), response backchannels are chosen preferentially in order to keep the user interested or to increase the interest of the user if it is the beginning of the interaction. After a while when the agent detects that the user is more and more disinterested (LUI < 0.4), the agent considers that the interaction is ending and stops progressively doing backchannels.

The selection is event-based and is done in real-time. Finally, the selection algorithm chooses the most appropriate backchannels based on the priority values according to the user's interest level and the context of the interaction.

5.4 Audiovisual backchannels

The backchannels are triggered and selected based on the communicative intentions that they represent, such as agreeing, liking, doubting, etc.. Each of these intentions can be realised in a range of ways; which behaviour is chosen for a given intention is decided in the behaviour planner component. Each SAL character has its own multimodal behaviour lexicon, which lists for every communicative intention the behaviours that can be used to realise that intention, as well as their respective probabilities of occurrence. These behaviours include visual-only behaviours such as a smile and a head nod, but they also contain audio-visual behaviours such as a smile, a head nod and a vocalisation such as *"right"* or *"myeah"* [16].

Audiovisual backchannels require lip synchronisation similar to speaking characters. Our solution consists of first generating the speech with timing information, using the same timing representation formats as for text-to-speech, and using the same viseme generation techniques as for fluent speech. This is suitable for vocalisations with a phonemic structure such as *"myeah"*, but is problematic for other vocalisations such as laughter, sighs, or a rapid intake of breath. In these cases, the viseme-based mouth shapes can only serve as coarse approximations of natural behaviour.

6. CONCLUSION

In this paper we presented the real-time platform we have developed within the EU project SEMAINE. From the input analysis and interpretation, we drive the behaviours of SAL agents. In particular we have presented a model of multimodal backchannels driven by audio input.

In the future, perceptive studies to evaluate the system will be performed by our SEMAINE partner [17].

A demo video of the system can be seen at: http://www.semaine-project.eu/

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

8. REFERENCES

[1] Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., and Stocksmeier, T. 2008. Modeling Embodied Feedback with Virtual Humans. Lecture Notes in Computer Science, 4930:18.

[2] Kopp, S., Jung, B., Lessmann, N., Wachsmuth, I. 2003. Max - A Multimodal Assistant in Virtual Reality Construction. KI-Küstliche Intelligenz 4/03, pp 11-17, Bremen: Verlag.

[3] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. 2007. Creating rapport with virtual agents. In et al., C. P., editor, 7th International Conference on Intelligent Virtual Agents, Paris, France.

[4] Vilhjálmsson, H. H. et al. 2007. The Behavior Markup Language: Recent developments and challenges. In Pelachaud, C., Martin, J.-C., Andr, E., Chollet, G., Karpouzis, K., and Pelé, D., editors. Proceedings of 7th International Conference on Intelligent Virtual Agents, volume 4722 of Lecture Notes in Computer Science, pages 99-111, Paris, France. Springer.

[5] Schröder, M. 2010. The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems. Advances in Human-Computer Interaction, vol. 2010, paper id 319406, doi:10.1155/2010/319406.

[6] Eyben, F., Wöllmer, M., Schuller, B. 2009. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009). IEEE, Amsterdam, The Netherlands, pp. 576-581.

[7] Schuller, B., et al. 2009. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. Image and Vision Computing Journal. Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior. ELSEVIER, Vol. 27, Issue 12, pp. 1760-1774.

[8] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A. 2009. Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, Merano, Italy.

[9] Maatman, R. M., Gratch, J., and Marsella, S. 2005. Natural behavior of a listening agent. In 5th International Conference on Interactive Virtual Agents. Kos, Greece.

[10] Ward, N. and Tsukahara, W. 2000. Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics, 23:1177-1207.

[11] ter Maat, M. and Heylen, D. 2009. Turn-management or Impression-management. In Intelligent Virtual Agents, 9th International Conference, IVA 2009. Z. M. Ruttkay, M. Kipp, A. Nijholt, and H.H. Vilhjálmsson (eds). Lecture Notes in Computer Science, volume 5773, Springer Verlag, Berlin, pp. 467-473.

[12] Chartrand, T., Maddux, W. and Lakin, J. Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. The new unconscious, pages 334-361, 2005.

[13] Bevacqua, E., Hyniewska, S. J. and Pelachaud, C. 2010. Positive influence of smile backchannels in ECAs. In Interacting with ECAs as Virtual Characters Workshop. The Ninth International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS'2010. Toronto, Canada.

[14] de Sevin, E. and Pelachaud, C. 2009. Real-time Backchannel Selection for ECAs according to User's Level of Interest. In Proceedings of Intelligent Virtual Agents 2009, IVA'09. Amsterdam, Holland.

[15] Goodwin, C. 1981. Conversational Organization: Interaction between Speakers and Hearers. Academic Press.

[16] Pammi, S. and Schröder, M. 2009. Annotating meaning of listener vocalizations for speech synthesis. In Proc. International Conference on Affective Computing & Intelligent Interaction. Amsterdam, The Netherlands: IEEE.

[17] Cowie, R. Perceiving emotion: towards a realistic understanding of the task Philosophical Transactions of the Royal Society B 364 (1535), 3515-3526