Multimodal Backchannels for Embodied Conversational Agents

Elisabetta Bevacqua[1], Sathish Pammi[2], Sylwia Julia Hyniewska[1], Marc Schröder[2], and Catherine Pelachaud[1]

[1] LTCI, CNRS - Telecom ParisTech, 37/39 rue Dareau, 75014 Paris, France
[2] DFKI GmbH - Language Technology Lab, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany

Abstract. One of the most desirable characteristics of an Embodied Conversational Agent (ECA) is the capability of interacting with users in a human-like manner. While listening to a user, an ECA should be able to provide backchannel signals through visual and acoustic modalities. In this work we propose an improvement of our previous system to generate multimodal backchannel signals on visual *and* acoustic modalities. A perceptual study has been performed to understand how context-free multimodal backchannels are interpreted by users.

1 Introduction

In the past twenty years several researchers in the human-machine interface field have concentrated their efforts in the development of Embodied Conversational Agents (ECAs): virtual humanoid entities able to interact with users in a human-like manner. To sustain a natural interaction with users, conversational agents must be able to exhibit appropriate behaviour while speaking and while listening. In this paper we focus on the listener's behaviour and in particular on the signals performed by the interlocutor. To describe this type of signals, Yngve [Yng70] introduced the term *backchannel*: non-intrusive acoustic and visual signals provided during the speaker's turn. According to Allwood et al. and Poggi [ANA93, Pog07], acoustic and visual backchannels provide information about the basic communicative functions, as perception, attention, interest, understanding, attitude (e.g., belief, liking) and acceptance towards what the speaker is saying. In previous works [HBTP07, BHTP07] we performed perceptual studies on unimodal backchannel signals displayed on visual modality. The results of these evaluations helped us to build up a library (called *backchannel lexicon*) of prototypical backchannel signals to be used in a listener module for an ECA. However, backchannels are provided not only through the visual modality, but also through voice by uttering paraverbals, words or short sentences [Gar98, ANA93]. In this work we propose to improve user-agent interaction by introducing multimodal signals in the backchannels performed by our ECA. Moreover, we present a perceptual study that we performed to get a better understanding about how multimodal backchannels are interpreted by users.

J. Allbeck et al. (Eds.): IVA 2010, LNAI 6356, pp. 194–200, 2010.

Such an evaluation allows us to extend the backchannel lexicon. This work is set within the Sensitive Artificial Listening Agent (SAL) project that is part of the EU project SEMAINE (http://www.semaine-project.eu). Such a project aims to build an autonomous talking agent able to exhibit appropriate behaviour while listening to a user. The agent has to encourage the user into talking pulling him towards specific emotional states.

The following Section provides an overview of the related works. In Section 3 we explain how visual and acoustic backchannels are generated. Section 4 describes our ECA system. Finally, we describe the perceptual study we have conducted and we analyse the results.

2 Related Works

Past researches on ECAs have provided first approaches to the implementation of a backchannel model. K. R. Thórisson [Thö96] developed a talking head, called Gandalf, capable of interacting with users using verbal and visual signals (like a short utterance or a head nod). REA, the Real Estate Agent developed by Cassell et al. [CB99], is able to provide backchannel signals such as paraverbals (e.g. *mmhmm*), head nods or a short statements (like *I see*). Its task consists in showing users the characteristics of houses displayed behind her. Gratch et al. [GWG$^+$07] developed the "Rapport Agent", an agent that provides solely visual backchannels when listening. The system analyzes the user's visual behaviour (nod, shake, head movement, mimicry) and some features of the user's voice to decide when backchannel must be triggered and which signal must be dispayed. Morency et al. [MdKG09] proposed an enhancement of this type of system introducing a machine learning method to find the speaker's multimodal features that are important and can affect timing of the agent backchannel. Kopp et al. [KAG$^+$08] proposed a backchannel model based on a reasoning and deliberative processing that plans how and when the agent must react according to its intentions, beliefs and desires.

All these models above take into account a small number of multimodal backchannel signals, moreover their communicative functions are not really defined. Through this work we aim to improve our system by introducing a large set of vocalizations to generate multimodal backchannels. Moreover we want to define the meaning that these signals convey when displayed by an ECA.

3 Multimodal Backchannels

Visual signals. As a first step we endowed our agent with the capability of providing visual backchannel signals while listening to a user. From the literature [ANA93, Pog07] we selected twelve frequent meanings related to the listener's reactions and we performed perceptual studies to understand how users associate these meanings to a set of visual backchannels displayed by a virtual agent [HBTP07, BHTP07]. The results of these evaluations allowed us to define some associations between the listener's communicative functions and a set of visual signals. Each of these associations represents one element of the agent's backchannel lexicon. Within the SEMAINE project new visual signals have been added. Since SAL provides four agents with different emotional traits, the backchannel lexicon has been expanded by introducing signals that are typical to each agent. For example, Spike, who is angry and aggressive, scowls even when it performs a head nod to show agreement.

Endowing TTS with vocal backchannels. Like visual backchannels, vocal backchannels also play an important role in communicating listener intentions while the interlocutor is talking. For the generation of vocal backchannels, an ECA should be able to use the same voice with which it speaks. As the SEMAINE project is already using expressive voices available in MARY TTS [ST03, SPT09], our work requires the addition of a new functionality to TTS: to generate vocal backchannels. To collect database of listener vocalisations as they appear natural only in conversation, in addition to speech synthesis recordings, free dialogue of around 30 minutes was recorded with a professional female British actor with whom we had recorded a happy expressive speech synthesis database. The actor was instructed to participate in a free dialogue, but to take predominantly a listener role. Listener vocalisations were marked on the time axis and transcribed as a single (pseudo-)word, such as *myeah* or *(laughter)*. The dialogue speech contains 174 spontaneous listener vocalisations from the actor. Among them, most frequent segmental forms are *yeah*, *(sigh)*, *(laughter)*, *mhmh*, *(gasp)*, *oh*. Phonetic alignment of speech is always required for ECA's lip synchronisation. Hand-labelled phonetic segment labels for all vocalisations were provided by a phonetically trained student assistant. The manual labels of a vocalisation contain time-stamps of each phonetic segment as well as corresponding suitable phone description. 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. Annotations of intonation, voice quality and meaning are also performed. The MARY TTS framework was extended to generate listener vocalisations based on an XML request. The TTS system stores the recorded audio of each vocalisation together with phone segment labels and features representing the segmental form, intonation, voice quality and possible meanings of the vocalisation, as annotated previously. At run-time synthesis, the selection of a vocalisation is an extension to the MARY TTS unit selection mechanism. A cost function which operates on the features of each vocalisation finds the most suitable vocalisation for a given markup.

4 System Overview

Our system is implemented on top of the SEMAINE API, a distributed multi-platform component integration framework for real-time interactive systems [Sch10]. The communication passes via the message-oriented middleware ActiveMQ. The architecture of our system is shown in Figure 1. Components (that

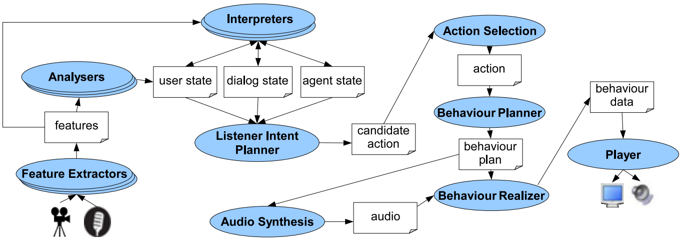

Fig. 1. Architecture of the SAL system

receive and send messages) are shown as ovals, message types as white rectangles. The architecture can generate the agent's behaviour both while it speaks and it listens, however in this paper we are interested in the generation of the listener behaviour. 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). The analyzers components derive some sense from the raw features in a context free manner; then the interpreters derive the system's *current best guess* regarding the state of the user and the dialogue. In parallel, the **Listener Intent Planner**, can trigger backchannels according to the user's behaviour, which are filtered by an Action Selection. Then, the **Behaviour Planner** computes a list of adequate visual behavioural signals for each communicative function the agent aims to transmit through the backchannel. The acoustic signals are generated by the **Audio Synthesis**. This module uses MARY TTS (see Section 3). MARY TTS looks up available vocalisations for the given speaker and will generate the most appropriate vocalisation found for the request. Finally, the agent behaviour is realized by the **Behaviour Realizer** module and rendered by a 3D character player.

5 Evaluation Description

We performed an evaluation study to analyze multimodal backchannels. To this purpose, we asked subjects to judge a set of multimodal signals performed by the 3D agent Greta [NBMP09]. Like in our previous studies, we considered in this perceptual evaluation the twelve meanings: *agreement, disagreement, acceptance, refusal, interest, not interest, belief, disbelief, understanding, not understanding, liking, disliking.* The signals were context-free, that is without knowing the discursive context of the speaker's speech. To create videos we selected 7 visual signals and 8 audio signals (7 vocalisations plus silence). The visual signals were chosen among those we studied in previous evaluations [HBTP07, BHTP07]. The vocalisations were selected using an informal listening test. Initially, three participants assigned each of the 174 vocalisations produced by the speaker to one of the 12 meanings used in this experiment. We then selected the seven stimuli which seemed least ambiguous for their respective meaning, in order to cover a reasonable range of different vocalisations. We generated 56 multimodal signals as the combinations of the visual and acoustic cues selected. Since there was quite a lot of videos to evaluate, we decided to split them in three sets (A, B and C). We hypothesized that:

- **Hp1:** the strongest attribution of a meaning will be conveyed by the multimodal signals obtained by the combination of visual and acoustic cues representative of the given meaning.
- **Hp2:** in some occasion, multimodal signals convey a meaning different from the ones associated to the particular visual and acoustic cues when presented on their own.
- **Hp3:** visual and acoustic signals that have strongly opposite meanings are rated as nonsense: like *nod+no*, *shake+ok*, *shake+yeah*.

55 participants (22 women, 33 men) with a mean age of 31.5 years, mainly from France (33%), Italy (18%), accessed anonymously to the evaluation through a web browser. The first page provided instructions, the second collected demographic information. Then the multimodal signals were played one at a time. Participants used a bipolar 7-points Likert scale: from -3 (extremely negative attribution) to +3 (extremely positive attribution). The evaluation was in English.

Table 1. Meanings significantly associated to the multimodal backchannels. AG=agreement, AC=acceptance, DA=disagreement, R=refusal, L=liking, NL=no liking, B=belief, DB=disbelief, I=interest, NI=no interest, U=understanding, NU=no understanding

	ok	**ooh**	**gosh**	**really**	**yeah**	**no**	**m-mh**	**(silence)**
raise eyebrows	AG, AC, U	U		I		NL		
nod	B, AG, AC, U	AC, L, U, I	AG, AC, U, I	L, U, I	B, AG, I AC, U		B, AC, AG	B, AG, AC, U
smile	B, AG, AC, U, L, I	B, AG, AC, U, L, I	AG, L	AG, AC, U, L, I	AG, AC, U, L, I	DB	B, AG, AC, L	
frown	AG, AC	NL	NL	NL, I		DA, NL		DB, N, U, NL
raise left eyebrow	AG, AC	U	DB	DB, I		DB, R		DB, NL
shake		DB, NL	DB, NL	DB, NI	DB, NL	DA, R	DA, R, NL	DA, DB, R, NL, NI
tilt&frown	AC	U		DB, I	AC, L	DA, R, NL		DB, NU

5.1 Results and Discussion

The 95% confidence interval was calculated for all the meanings. Table 1 reports all signals for which the mean was significantly above zero (for positive meanings) or below zero (for negative meanings). For each dimension of meaning (i.e. agreement/disagreement, acceptance/refusal, etc.) we performed a repeated measures ANOVA. We obtained that for all dimensions there was an effect of different visual cues ($p<.05$) and an effect of acoustic cues ($p<.05$). We did not find any effect of the interaction between the visual and acoustic cues (p>.05). Some t-test results are reported here detail. The signal *nod+yeah* (N=12, mean=2.75) was more strongly judged as showing agreement than any other signal (p<.05). *Nod* (N=12, mean=2.07) has the second highest attribution of agreement. The signal *shake+no* (N=14, mean=-1.71) was not more strongly judged as showing disagreement than the other signals. The highest disagreement mean is for *shake* (N=14, mean=-2.07), however it is not significantly different from *shake+no*, *shake+m-mh* (p>.05). There is a difference between *shake* and *shake+yeah*, which is the fourth highest disagreement attribution (0.18). The signal *raise eyebrows+gosh* was not even significantly associated to interest. The highest meaning of interest was equally attributed to *smile+ok*, *nod+ok*, *nod+ooh*, *smile+ooh* (p>.05). Highest attribution of understanding was observed for *raise eyebrows+ooh*, *nod+ooh*, *nod+really*, *nod+yeah* and *nod. Raise eyebrows+ooh* (mean=1.56) was not more strongly judged as showing agreement than the other signals. A significant difference was even found between *nod-ooh* and *raise eyebrows+ooh* (p<.05): *nod-ooh* was more strongly associated to the understanding than *raise eyebrows+ooh.* In conclusion our first hypothesis has been only partially satisfied. As regard to the third hypothesis we saw that four multimodal signals were significantly rated as nonsense: *nod+no* (p<.05), *shake+yeah* (p<.05), *shake+ok* (p<.05) and *shake+really* (p<.05).

Our first hypothesis has been only partially satisfied. Results showed that the strongest attribution for a meaning is not always conveyed by the multimodal signals obtained by the combination of visual and acoustic cues representative of the given meaning. That means that the meaning conveyed by a multimodal backchannel cannot be simply inferred by the meaning of each visual and acoustic cues that compose it. It must be considered and studied as a whole to determine the meaning it transmits when displayed by virtual agents. Moreover, we found that some multimodal signals convey a meaning different from the ones associated to the particular visual and acoustic cues when presented on their own (Hp2). Our evaluation showed also that multimodal signals composed by visual and acoustic cues that have strongly opposite meanings are rated as nonsense. As expected *nod+no*, *shake+yeah*, *shake+ok* and *shake+really* were rated as senseless. What is more, a high attribution of nonsense does not necessarily exclude the attribution of other meanings. Thus, the high nonsense signal of *shake+yeah* was also highly judged as showing disbelief. A possible explanation would be that these signals might be particularly context depend. This evaluation gave us a better insight about several multimodal backchannels and the meaning they convey. The results have been used to enrich and expand the backchannel lexicon of our virtual agent.

6 Conclusion

We have presented an ECA system able to generate a wide variety of multimodal backchannel signals simulating listening behaviour. A perceptual study has been conducted in order to understand how context-free multimodal backchannels are interpreted by users.

Acknowledgement

This work has been funded by the STREP SEMAINE project IST-211486 (http://www.semaine-project.eu).

References

[ANA93] Allwood, J., Nivre, J., Ahlsn, E.: On the semantics and pragmatics of linguistic feedback. Semantics 9(1) (1993)

[BHTP07] Bevacqua, E., Heylen, D., Tellier, M., Pelachaud, C.: Facial feedback signals for ECAs. In: AISB 2007 Annual convention, workshop "Mindful Environments", Newcastle upon Tyne, UK, pp. 147–153 (April 2007)

[CB99] Cassell, J., Bickmore, T.: Embodiment in conversational interfaces: Rean Human Factors in Computing Systems, Pittsburgh, PA (1999)

[Gar98] Gardner, R.: Between Speaking and Listening: The Vocalisation of Understandings. Applied Linguistics 19(2), 204–224 (1998)

[GWG$^+$07] Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D., et al. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)

[HBTP07] Heylen, D., Bevacqua, E., Tellier, M., Pelachaud, C.: Searching for prototypical facial feedback signals. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 147–153. Springer, Heidelberg (2007)

[KAG$^+$08] Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T.: Modeling embodied feedback with virtual humans. In: Wachsmuth, I., Knoblich, G. (eds.) ZiF Research Group International Workshop. LNCS (LNAI), vol. 4930, pp. 18–37. Springer, Heidelberg (2008)

[MdKG09] Morency, L.-P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. In: Autonomous Agents and Multi-Agent Systems (2009)

[NBMP09] Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C.: Greta: an interactive expressive eca system. In: AAMAS 2009 - Autonomous Agents and MultiAgent Systems, Budapest, Hungary (2009)

[Pog07] Poggi, I.: Mind, hands, face and body. A goal and belief view of multimodal communication. Weidler, Berlin (2007)

[Sch10] Schröder, M.: The semaine api: Towards a standards-based framework for building emotion-oriented systems. In: Advances in Human-Computer Interaction (2010)

[SPT09] Schröder, M., Pammi, S., Türk, O.: Multilingual mary tts participation in the blizzard challenge 2009. In: Proc. Blizzard Challenge 2009 (2009)

[ST03] Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech technology 6, 365–377 (2003)

[Thö96] Thórisson, K.R.: Communiative Humanoids: A Computational Model of Psychosocial Dialogue Skills. PhD thesis, MIT Media Laboratory (1996)

[Yng70] Yngve, V.: On getting a word in edgewise. In: Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, pp. 567–577 (1970)