Towards Influencing of the Conversational Agent Mental State in the Task of Active Listening

Stanislav Ondáš[1], Elisabetta Bevacqua[2], Jozef Juhár[1], and Peter Demeter[1]

[1] Technical University of Košice, Faculty of Electrical Engineering and Informatics,
Laboratory of speech technologies, Letná 9, Košice, Slovakia
[2] CNRS - Telecom ParisTech, 37/39 rue Dareau,
75014 Paris, France
{stanislav.ondas,jozef.juhar}@tuke.sk,
elisabetta.bevacqua@telecom-paristech.fr, demeter.peto@gmail.com

Abstract. The proposed paper describes an approach that was used to influence conversational agent Greta's mental state. The beginning this paper introduces the problem of conversational agents, especially in the listener role. The listener's backchannels also influence its mental state. The simple agent state manager was developed to impact Greta's internal state. After describing this manager, we present an overview of evaluation experiments carried out to obtain information about agent state manager functionality, as well as the impact of the mental state changes on the overall interaction.

Keywords: Embodied conversational agent, listener, mental state, backchannel.

1 Introduction

Researches have shown that people tend to interact with computers characterized by human-like attributes as if they were really humans [1], [24]. Consequently, the more humane-machine interfaces are consistent with human style of communication, the more their use becomes easy and accessible [2]. Such level of consistency could be reached using Embodied Conversational Agents (ECAs): computer-generated animated characters that are able to carry on natural, human-like communication with users [3]. In this work, to perform our tests, we use the Embodied Conversational Agent Greta [4] developed at the Telecom ParisTech Research Centre, see Section 3.

Human interlocutors in the dialog take the roles of the speaker and the listener. Each of these roles has their own behavioral properties. There is also an effort to give these properties to the virtual humans. Earlier projects usually deal with the behavior of the speaker. Presented work is focused on the modeling of the behavior of virtual listener. Active listening is an important dimension of the conversational agent behavior. Backchannel produced by agents in this role relates to three main factors – what was seen, what was heard and the internal mental state of the listener. Producing of backchannel signals require an understanding of interlocutor's speech. Thus the key limitation relates to the unavailability of partial interpretation of speaker's utterance until he is finished and consecutively to the problems with the synchronization [5].

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNCS 5967, pp. 113–121, 2010.

The proposed work has their background in the Project 7: “Multimodal Feedback from Robots and Agents in a Storytelling Experiment”, which was solved on eNTERFACE’08 Summer School [6]. This project lies at the intersection between Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) and was especially focused on active listening and generating feedback during storytelling task. [7].One of the goals of this project was to give the properties of active listener to the both the embodied conversational agent Greta [4] and the robotic dog Aibo.

The proposed paper mainly presents an approach, which was used for influencing of the internal mental state of the listener. For this purpose was developed a simple Agent State Manager (ASM), which uses speech recognition for obtaining information about what was spoken. The second part of the paper deals with the evaluation experiment, which was carried out with 50 students, who told a story to the ECA Greta. Then they filled out the questionnaires about their impression from the interaction.

2 Background – Listener's Behavior

During a conversation the listener is called to provide information on the successfulness of the communication. Through verbal and non verbal signals, called “backchannels” [8], [26], a listener can show his level of engagement in the conversation, providing information about the basic communicative functions, as perception, attention, interest, understanding, attitude (e.g., belief, liking and so on) and acceptance towards what the speaker is saying [9], [23]. Backchannels can be the result of two evaluation stages of the reception of the speaker's message [10]: at the first stage, a non-conscious appraisal of the perceived information can generate automatic behavior, for example to show lack of contact or perception. At a second stage, the more aware evaluation, involving memory, understanding and other cognitive processes, can generate signals to show understanding and other attitudinal reactions, like acceptance or rejection, belief or disbelief and so on. We call *reactive* the behavior derived from perception processing and response the more aware behavior generated by cognitive processing. Another particular form of backchannel is the *mimicry* [12] of the speaker's behavior. By mimicry we mean the behavior displayed by an individual who does what another person does [11]. This type of behavior has been proven to play quite an important role during conversations. When present, it makes the conversation run more smoothly [12], and helps to regulate the conversational flow. For example, listeners often mirror speaker's postural shifts at the end of a discourse segment and this helps the exchange of speaking turns [13].

3 GRETA: An Embodied Conversational Agent

Greta is an Embodied Conversational Agent that communicates through several modalities (like head, face, gaze, gesture and torso) while interacting with a user. The agent architecture follows the design methodology proposed in [14] and is compatible with the standard SAIBA framework (Situation, Agent, Intention, Behavior, Animation) [15].

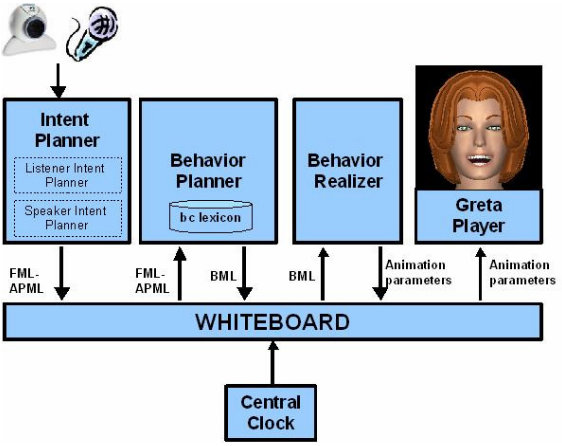

Fig. 1. Greta's Architecture

The architecture of the agent Greta (Figure 2) is modular and distributed. Each module exchanges information and data through a central message system. We use the concept of whiteboard [14] that allows internal modules and external software to be integrated easily. The Intent Planner module is divided into two sub-modules, the Listener Intent Planner and the Speaker Intent Planner that decide of the agent's communicative intentions respectively in the role of the listener and in the role of the speaker. The Intent Planner encodes the agent's communicative intentions into the FML-APML language (Function Markup Language - Affective Presentation Markup Language), a first approach to the FML language (Function Markup Language). The Behavior Planner module receives as input the agent's communicative intentions written in FML-APML and, to convey them, it schedules a number of communicative signals (e.g., speech, facial expressions, gestures) which are encoded with the Behavior Markup Language (BML). Such a language specifies the verbal and non-verbal behaviors of ECAs [15]. The task of the Behavior Realizer is to realize the behaviors scheduled by the Behavior Planner. Finally, the animation is played in the Greta Player. The synchronization of all modules in the distributed environment is ensured by the Central Clock.

3.1 Generation of the Listener's Behavior

The Listener Intent Planner module is in charge of the computation of the agent's behaviors while being a listener when conversing with a user. This component encompasses two modules called response/reactive backchannel and mimicry. Research has shown that there is a strong correlation between backchannel signals and the verbal and nonverbal behaviors performed by the speaker [5], [17]. From the literature [5], [17] we have fixed some probabilistic rules to decide when a backchannel signal should be triggered. Our system analyzes speaker's behaviors looking for those 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. Then, the response/reactive backchannel, and mimicry modules compute which type of backchannel should be displayed. The response/reactive backchannel module uses information about the agent's beliefs towards the speaker's speech to calculate the response backchannel signal. We use Allwood's and Poggi's taxonomies of communicative functions of backchannels [9]: understanding and attitudinal reactions (liking, accepting, agreeing, believing, being interested).

A lexicon of backchannels has been elaborated [18], [23]. The response module selects which signals to display from the lexicon depending on the agent's reaction towards the speaker's speech. When no information about the agent's beliefs towards the speaker's speech is given, the response/reactive module selects a pre-defined backchannel among those signals that have been proven to show contact and perception, like head nod and raise eyebrows. When fully engaged in an interaction, mimicry of behaviors between interactants may happen [19]. The mimicry module determines which signals would mimic the agent. A selection algorithm determines which backchannels to display among all the potential signals that are outputted by the two modules.

4 Agent’s Mental State Influencing

As was said above the response/reactive backchannel module requires information about *agent’s beliefs* and such information can be given predominantly from speaker’s speech. Mentioned communicative functions of backchannels relate to agent’s belief and we can say that these functions represent the internal agent mental state.

The mental state of the agent Greta determines how the agent reacts to the user’s speech and it is represented through mentioned functions: agreement, disagreement, acceptance, refusal, belief, disbelief, liking, disliking, interest, no interest, understanding, no understanding [20]. In the ECA Greta, the agent’s mental state has been static and tied together with the agent’s baseline. The agent’s baseline is defined in [21] as a set of numeric parameters that represents the agent’s behavior tendencies.

It is clear, that the mental state of the human interlocutors in the interaction is changing. These changes relates to large range of various factors as are the relationship of the interlocutors, the initial communication intentions and feelings, earlier interactions, the common knowledge, and so on. But the main (direct) impact on the mental state of participants in the interaction has the content of the interaction. Therefore, we need some degree of content understanding, if we want to change the agent’s mental state during the interaction. For these purposes the agent state manager (ASM) was developed.

4.1 Agent State Manager (ASM)

The key functionality of the ASM is to change agent’s mental state according to the speaker’s speech. It is necessary to extract partial meaning before the speaker stops speaking. The question was how can we do this? At the beginning we have selected the simplest way – to process speaker’s utterances over the words, because the words are the smallest meaningful parts of the utterance. The designed ASM is shown on Fig. 2.

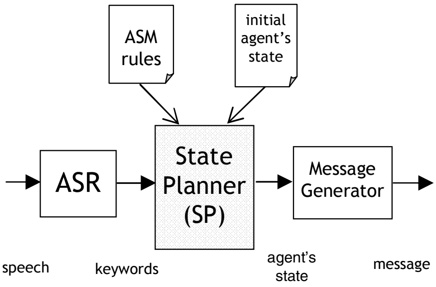

Fig. 2. Architecture of the agent state manager

It consists of three main parts – Automatic Speech Recognition engine (ASR), State planner (SP) and Message generator. The ASM takes words, which are spoken by the speaker and it impacts the mental state of the agent according to the set of rules. At the beginning the initial agent state is loaded from an XML file. The message generator produces the messages, which informs Greta about the new mental state.

The speech recognition engine we have used is based on ATK/HTK [25]. It supports Slovak language and it works in keyword-spotting mode. The ASR engine uses triphone acoustic models, which were trained on Slovak SpeechDat-E database [21], [22]. As a language model we have used a speech grammar which enables the recognition of keywords in phrases. Other words are modeled as "filler words".

The State Planner is the main part of the agent state manager. Its task is to modify the state of the agent according to spoken input (keywords). For this purpose it uses a rule-based approach. This component is initialized with the initial agent state as well as with a set of rules loaded from the file. Each rule consists of the following items: *feature*, *keyword*, *step*, *max_value*, *opposite_feature*, *action*. The items *opposite_feature* and *action* are optional. When the speech recognizer recognizes a keyword, the State Planner looks for an appropriate rule. If ASM finds such rule, it increases the value of the related feature by a given *step* value, while *max_value* is not reached. If some *opposite_feature* is defined in the rule, it is decreased by the same *step*. The State Planner can also create a requirement for some action as are headnod, smile or headshake, if in appropriate rule is defined the parameter *action*.

The last part of the ASM is the Message Generator, which is responsible for preparing an XML file containing the values (features), which represent the new mental state of the agent. It sends the XML file through a TCP socket to the ECA Greta.

5 The Storytelling Experiment

The main goal of the storytelling experiment, which we carried out, was to assess the functionality of the developed agent state manager and to observe the impact of the agent's mental state changes on the human speaker during his storytelling to the agent. The cartoon about Silvester and Tweety was selected as the scenario of this storytelling. We assumed that after such funny story, levels of liking, interest, understanding and acceptance will increase and the agent will produce appropriate (richer) backchannel signals. The backchannel produced by Greta depended only on the speaker's speech. Thus means that other factors did not influenced the interaction.

There were more than 50 speakers, who told the story to the ECA Greta. Before the experiment were necessary to define the keywords for ASR engine, to create profiles (agent's baseline and a set of rules for ASM), to design evaluation questionnaires and to establish the place of the experiment. For the purpose of evaluation were used both the questionnaires and the log files produced by ASM.

5.1 Setup of the Experiment

At the beginning we needed to select keywords for speech recognition. Therefore ten storytelling of mentioned cartoon were recorded. Then the recordings were analyzed in terms of word counts. After eliminating conjunctions and pronouns, which were naturally the most frequent words, we identified a group of keywords: vtáčik (bird), kocúr (tomcat), Silvester, babka (grandmother), kufre (baggages), klietka (cage), pani (lady), poslíček (callboy), dáždnik (umbrella), hlava (head).

We wanted also to compare different profiles of the agent and to assess speaker's impression from the interaction with them. These profiles determine how the agent generates backchannel during the storytelling. Three kinds of profiles – Spike, Poppy and Ann were created, which consist of different rule files for configuration of ASM, different initial mental states and different agent's baselines (adopted from [20]). Accordingly, speakers were divided in to three groups, in such manner that each group told the story to the one of the ECA's profile.

The storytelling experiment was carried out in our laboratory. The ECA ran on the PC with soundcard and the speaker sat in the front of the monitor and told the story in to the headset with microphone. After his storytelling, he filled out the questionnaire about his impression of realized interaction.

5.2 The Results of the Subjective Evaluation

The subjective evaluation was based on questionnaires fulfilling. For this purpose the questionnaire with five questions was designed. The first question was:

1. "*Did you have the feeling that Agent's mood was changing during your storytelling?*" - 90% respondents, who told the story to the Poppy said "yes". The same response was selected also by speakers who communicated with Ann (about 83%) and Spike (about 67%). It means that changing of agent's state influenced the perception of listener's mood by speaker, what relates to the fact, that ECA showed more liking and interest (see also Fig.3).

2. "*Did you have the feeling, that Agent understood you?*" - The most of the speakers (about 92%) who told the story to Ann and 70% of speakers who told the story to Poppy, had that feeling. In the Spike profile, it was 50% of speakers. These results matched also to the level of understanding on Fig. 3. We can say that the internal state of the agent was communicated appropriately to the speaker during the backchannel.

3. "*How did Agent like your storytelling?*" – It was Ann who liked the stories the most. (58.3%) Spike disliked the stories the most (66.7%) and Poppy disliked the stories as well (50%). The lower values of "liking" in the case of Spike and Poppy do

not match the “liking” level. There are two factors, which could cause this situation – the zero initial value of “liking” parameter and not good rules for ASM, which impact “liking” parameter.

4. “*Was Agent interested in your storytelling?*” - The most interested “person” was Ann (the answer “mostly yes” – 50%). On the contrary, agents Spike and Poppy were “mostly not” interested in the storytelling (33.3% and 50%). This is the similar situation as was in the previous question.

5. “*How much did Agent behave like a human being?*” - All speakers who were communicating with Spike labeled “mostly not” answer. In the profile Poppy the answers “mostly not” and “mostly yes” were equal (40%). The respondents had the feeling that the agent in the profile Ann mostly behaves like a human being (58.3%). We supposed such results, because the profile Ann has higher initial values of “acceptance”, “belief” and “liking” parameters as well as different set of ASM rules, which allow rapid increasing of mental state parameters.

5.3 The Results of the Objective Evaluation

The objective part of the evaluation experiment was focused on comparing initial agent state and the mental state after the storytelling. Figure 3 brings a comparison of those values for all profiles. We have used log files produced by ASM for obtaining these values.

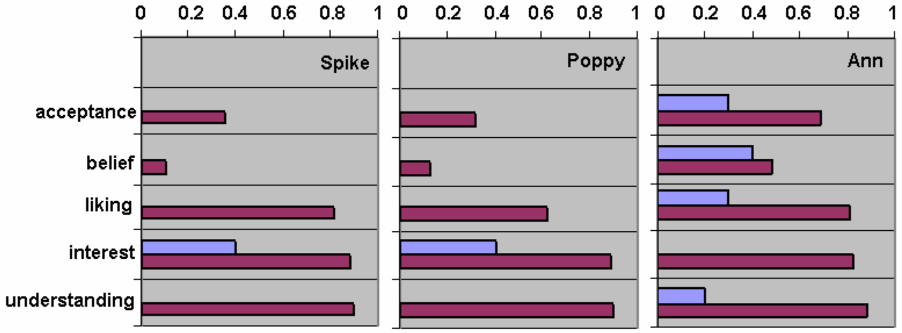

Fig. 3. The initial (blue) and final (violet) ECA's mental state

The first important fact, which we were observed is, that the ASM really impact the agent mental state. The second observation is that the values, which represent the degree of acceptance, belief, liking, interest and understanding, were increased. We have anticipated such situation because the story that was told was positive and funny. The highest values were reached in profile Ann. This results matches also to the results obtained from questionnaires. Speakers, who told the story to the ECA in profile Ann had the feeling, that ECA understood them (about 92%), that she was interested in this storytelling (50%) and she liked the story (about 58%).

6 Conclusions

Presented system is the first step towards the system for maintaining of agent's mental state during the conversation with human. Realized experiment shows that such approach is able to change mental state of the ECA in the role of the listener, but it is necessary to set up the initial agent's state and the rules for ASM carefully. The experiment also brings a need of some internal logic, which for example automatically decreases the value of interest in the case, when long interval of silence occurs. The next problem is that the value of "understanding" increases very quickly when a lot of keywords occur in the speaker's utterance.

Such system should integrate information from several modalities and it also should deliberate overall perception as well as information from memory. The statistical modeling is one of the possible ways.

In the future we want to include other inputs (modalities) in to the ASM – the beginning and the end of speech detection, a type of sentence recognition (question, statement…) based on prosody, and so on.

Acknowledgments. The work presented in this paper was supported by the Slovak Research and Development Agency under research project APVV-0369-07 and Ministry of Education of Slovak Republic under research projects AV 4/0006/07 and VEGA 1/4054/07. We would like to say special thanks to people from project No.7: "Multimodal Communication with Robots and Virtual Agents" of the eNTERFACE'08 summer school.

References

[1] Nass, C.I., et al.: Computers are social actors: a review of current research, pp. 137–162 (1997)

[2] Ball, G., Breese, J.: Emotion and personality in a conversational agent. Embodied Conversational Characters. MIT Press, Cambridge (2000)

[3] Cassell, J., Bickmcre, T., Campbell, L.: Designing Embodied Conversational Agents. Embodied Conversational Agents (2000)

[4] Pelachaud, C.: Multimodal expressive embodied conversational agents. In: MULTIMEDIA 2005: Proceedings of the 13th annual ACM international conference on Multimedia, pp. 683–689. ACM, New York (2005)

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

[6] eNTERFACE Summer School web page, `http://enterface08.limsi.fr/` (June 10, 2009)

[7] Al Moubayed, S., et al.: Multimodal Feedback from Robots and Agents in a Storytelling Experiment. In: Project 7: Final Project Report, eNTERFACE 2008, Paris, France, August 4-29 (2008)

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

[9] Allwood, J., et al.: On the semantics and pragmatics of linguistic feedback. Semantics (1993)

[10] Kopp, S., et al.: 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)

[11] Van baaren, R.B. (ed.): Mimicry: a social perspective (February 10, 2003), http://webdoc.ubn.kun.nl/mono/b/baarenrvan/mimi.pdf

[12] Chartrand, T., Bargh, J.: The Chameleon Effect: The Perception-Behavior Link and Social Interaction. Personality and Social Psychology 76, 893–910 (1999)

[13] Cassell, J., et al.: Non-verbal cues for discourse structure. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics Morristown, NJ, USA, pp. 114–123 (2001)

[14] Thórisson, K.R., et al.: Whiteboards: Scheduling blackboards for semantic routing of messages & streams. In: AAAI 2005 Workshop on Modular Construction of Human-Like Intelligence, pp. 8–15 (2005)

[15] Vilhjálmsson, H.H., et al.: The Behavior Markup Language: Recent developments and challenges. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 99–111. Springer, Heidelberg (2007)

[16] Heylen, D., et al.: Why conversational agents do what they do? Functional representations for generating conversational agent behavior. In: The first Functional Markup Language workshop. The Seventh International Conference on Autonomous Agents and Multiagent Systems Estoril, Portugal (2008)

[17] Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in english and japanese. Journal of Pragmatics 23, 1177–1207 (2000)

[18] Bevacqua, E., et al.: Facial feedback signals for ECAs. In: AISB 2007 Annual convention, workshop “Mindful Environments”, Newcastle upon Tyne, UK, pp. 147–153 (2007)

[19] Lakin, J.L., et al.: Chameleon effect as social glue: Evidence for the evolutionary significance of nonconsious mimicry. Nonverbal Behavior 27(3), 145–162 (2003)

[20] Bevacqua, et al: A listening agent exhibiting variable behaviour. In: IVA 2008. LNCS (LNAI), vol. IVA 2008, pp. 262–269. Springer, Heidelberg (2008)

[21] Pollak, P., Cernocky, J., Boudy, J., Choukri, K., Rusko, M., Trnka, M., et al.: SpeechDat(E) Eastern European Telephone Speech Databases. In: Proc. LREC 2000 Satellite workshop XLDB - Very large Telephone Speech Databases, Athens, Greece, May 2000, pp. 20–25 (2000)

[22] Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G.: A noise robust multilingual reference recognizer based on SpeechDat (II). In: Proc. ICSLP 2000, Beijing, China (2000)

[23] Heylen, D., et al.: 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)

[24] Reeves, B., Nass, C.: The media equation: How people treat computers, television and new media like real people and places (1996)

[25] Young, S.: ATK: An application Toolkit for HTK, version 1.6. Cambridge University, Cambridge (2007)

[26] Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn taking for conversation. Language 50(4), 696–735 (1974)