EMPHASIZING ON THE TIMING AND TYPE
Enhancing the Backchannel Performance of Virtual Agent

Xia Mao, Na Luo and Yuli Xue
School of Electronic and Information Engineering, Beihang University, Bejing, China

Keywords: Human-Computer Interaction (HCI), Virtual Agent, Backchannel, Personality Rules, Emotional Backchannel Lexicon.

Abstract: Addressing backchannel feedbacks to virtual agent listener gives the agent human-like conversation skills and creates rapport in Human-Computer Interaction. We argue the limitations of current approaches in predicting and generating backchannel. Following two hypotheses emphasizing on the timing and type of backchannel, we introduce an improved system to enhance the agent listener's performance. By using Newcastle Personality Assessor before parasocial consensus sampling and then neural networks, we can obtain the personality rules and select different backchannel timing thresholds for specific agent listener according to its own personality. After a context-free perceptual study, we will build two emotional backchannel lexicons showing positive affection and negative affection respectively. In accordance with the empathy strategy, the system will select one type of backchannel from the corresponding emotional backchannel lexicon. The improved system will be more suitable for different conversation occasions and greatly increase the naturalness between the human speaker and the virtual agent listener in the future.

1 INTRODUCTION

Human-Computer Interaction focuses on human's feelings and aims at making the interaction process more natural and efficient. In multimodal interactions, users can use their facial expressions, voice, gestures, postures to express themselves to the computer, but corresponding feedbacks from the computer are limited. Virtual agent, which is the geometrical and active presentation of human in virtual environment and can realize multimodal interaction with people, has been widely applied. At the very start, virtual agents used in human-computer interaction were acting passively and stiffly. With the development of technology, users prefer more intelligent, autonomous, interactive, reactive agents to ones with predefined behaviour. Consequently, researches on how to make agents percept and act like human and how to make the interaction between agent and human more natural and vivid have won universal attentions.

When two people have a conversation, the listener will naturely produce some behaviours or short utterances while the speaker is talking. These feedbacks include nod, smile, shake head, say 'yeah', 'hmm', etc. However, the listener is usually not conscious of giving these feedbacks. We deem that it is related to the listener's subconsciousness.

Backchannel is pervasive in conversations and is an important kind of feedback in a dialogue as a signal of presenting listener's interest and encouraging speaker to continue speaking. It was first found by Victor Yngve (1970). He found backchannel communication when analyzing English conversation. Later, Duncan (1974) and his colleagues (Duncan and Fiske, 1977) termed this listener's feedback phenomenon as backchannel. Earlier study found that when people interacted with others, they used backchannel feedbacks such as speech prosody, gesture, gaze, posture and facial expression to establish a sense of rapport. Backchannel plays an important role in everyday conversation, especially in the speaker-listener dialogue. Application of appropriate backchannel has been found to improve the narrator's performance (Bavelas et al., 2000).

In face-to-face interaction between a human speaker and a virtual agent listener, addressing backchannel feedbacks to virtual agent listener gives the agent human-like conversation skills and creates rapport in Human-Computer Interaction. Researchers have made great efforts in predicting backchannel for the agent listener. Ward and

Mao X., Luo N. and Xue Y..
EMPHASIZING ON THE TIMING AND TYPE - Enhancing the Backchannel Performance of Virtual Agent.
DOI: 10.5220/0003833102590263
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 259-263
ISBN: 978-989-8425-96-6

Tsukahara (2000) paid attention to regions of low pitch late in an utterance and made five rules to produce backchannel feedback by the low pitch cue. Cathcart et al. (2003) proposed that backchannel often produced after a short pause in the speaker's discourse. In addition to only audio-based prediction, visual evidences are also very useful in the multimodal interaction. Kendon (1967) and Bavelas et al. (2002) found out that there were relationships between eye gaze and backchannels. Maatman et al. (2005) presented a mapping from posture shifts, head movements and speech quality to agent listening behaviours.

However, current backchannel systems are not well accepted in the subjective evaluation progress. The agent listener's feedbacks are considered not natural and precise. In order to enhance the backchannel performance of the agent listener, we propose to build an improved system which emphasizes on the timing and type of backchannel and is implemented by analyzing the listener's personality and emotion state.

2 CURRENT APPROACHES AND LIMITATIONS

A virtual agent with appropriate backchannel feedbacks do act more like a real human listener, but we find that the subjective evaluation results of the current experiments are not so satisfactory in rapport scale.

Poppe et al. (2010) evaluated six different multimodal rule-based strategies for backchannel generation in face-to-face conversation. Features they used were speaker's speeches and eye gaze while backchannel performed by the agent listener was nods, vocalizations and the combination of them randomly. Evaluation results showed that Copy strategy got evident higher scores than the other five strategies. In Copy strategy, the timing of backchannel feedbacks was the same to the actual human listener. The other five strategies used by Poppe et al. almost covered all the rule-based approaches widely used for backchannel prediction, but the results of their subjective evaluations were not so good. When participants were asked how likely the agent's backchannel feedbacks were performed by a human listener, the average scores were all below 45(the full score is 100). It proved that current rule-based strategies themselves could difficultly perform as good as a human listener.

In contrast with rule-based methods, prediction models were wildly used in generating agent's backchannel. Morency et al. (2008) used sequential probabilistic model (Hidden Markov Model and Conditional Random Fields) to predict listener backchannel. Features they used were speaker's prosody, spoken words and eye gaze. Prediction results proved that their method with Conditional Random Fields outperformed Ward and Tsukahara's rule-based approach, but the value of precision and recall was 0.1862 and 0.4106 separately which was quiet low. It was a little far from the ideal backchannel prediction of human listener. Some of the reasons were due to the database which the prediction models learned from was limited and not taking into account individual differences.

Huang et al. (2010b) proposed a method to learn an effective prediction model from parasocial consensus sampling. This novel data collection method could collect large amount of behaviour data quickly and get rid of individual differences limitations in backchannel responses. Evaluation results showed that the virtual agent driven by Conditional Random Fields model trained on parasocial consensus sampling data obtained more rapport scale, perceived accuracy and naturalness than rule-based Rapport Agent (Gratch et al., 2006). Moreover, it performed better than the ones driven by actual human listener in low-rapport videos.

Huang's efforts in innovating methods make sense indeed. We deem that it is really necessary to research on backchannel feedback for virtual agent and develop improved approaches to enhance the agent's performance. We notice that the timing and type of backchannel are most significant to the human-like backchannel behaviour of virtual agent (Poppe et al., 2010), but in Huang et al.'s experiments, the agent listener only used nods as backchannel feedbacks. Furthermore, the listener's personality and emotion state were not taken into consideration during the backchannel predicting and generating process.

3 AN IMPROVED BACKCHANNEL SYSTEM

3.1 Two Hypotheses for the Backchannel System

Trait models of personality assume that traits influence behaviour, and that they are fundamental properties of an individual (McRorie et al., 2009). For improving rapport in the conversation, we also consider giving the agent capacity of empathy. In this way, the emotional state of the speaker will

influence the agent listener's feedback. Emphasizing on the timing and type of backchannel, we formulate two hypotheses:

H1: the timing of backchannel is related with the listener's personality.

H2: the type of backchannel is connected with the speaker's emotional state.

Following our two hypotheses, we propose to make an improvement in the previous backchannel system for virtual agent. Architecture of the improved system is illustrated in Figure 1. It consists of two main parts: (1) backchannel prediction, which extracts useful features from human speaker's video and predicts the timing of backchannel feedbacks; (2) backchannel generation, which generates action commands for the virtual agent listener and animates its backchannel feedback behaviours.

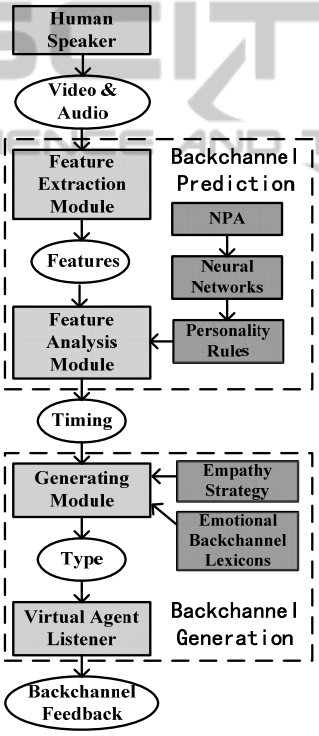

Figure 1: Architecture of the improved backchannel system for virtual agent system.

3.2 Backchannel Prediction

After recording conversational videos, we will apply parasocial consensus sampling (Huang et al., 2010a) to collect different listener's backchannel behaviour data and then learn a probabilistic model for backchannel prediction. According to our H1 Hypothesis, we plan to add listener's personality rules module to the prediction model which will make an influence on the timing of backchannel.

To find the relationship between the timing of backchannel and the listener's personality, we can ask the participants to fill out the Daniel Nettle's Newcastle Personality Assessor (NPA) before participating in parasocial consensus sampling. Results of the questionnaire will quantify the tester's personality on five dimensions: Extraversion, Neuroticism, Conscientious, Agreeableness, and Openness. After the sampling, we use neural networks to learn from participant's five personality dimensions and the total number of their backchannel feedbacks. In this way, we can obtain the personality rules related to the number of backchannel.

By establishing the personality rules, we can easily select different thresholds of response level for specific personality. The thresholds are used to decide the timing of backchannel by filtering out the feedbacks whose probabilities are low (Huang et al., 2010a). To keep the agent listener's backchannel fit its personality, we prepare to make the number of backchannel from parasocial consensus data closest to that from the personality rules.

The personality rules module enables our agent listener to be capable of different backchannel frequency to show multiple personality and makes our backchannel system more suitable for different conversation occasions in the future.

3.3 Backchannel Generation

Recently, researches have shown that empathic virtual agent enhance human-computer interaction (Prendinger et al., 2005). In our improved backchannel system, we hope that the agent listener can express the same type of emotion to the speaker through its backchannel feedback. For developing empathy strategy, the speaker's emotional states are divided into positive affection and negative affection, thus the virtual agent listener should show positive or negative affection in order to be similar with the speaker's emotional state.

According to H2 Hypothesis, if we can detect the speaker's emotional state, the most important step is to find out how one's emotion influences the type of backchannel behaviour. A context-free perceptual study is introduced to understand how various types of backchannel feedbacks are interpreted by users. We intend to generate multimodal backchannel for the agent listener as the combinations of visual and acoustic behaviours such as smile, nod, frown, shake head, say 'yeah', and say 'hmm'. The participants are asked to evaluate all the backchannel behaviours to assess the emotion expressed by the virtual agent

positive affection or negative affection. After the evaluation, we can build two emotional backchannel lexicons. By detecting features of the speaker, the system can analyze emotional state and randomly select one type of backchannel from the similar emotional backchannel lexicon of agent listener in accordance with the empathy strategy.

With the empathy strategy and emotional backchannel lexicon, the virtual agent listener can 'feel' the speaker's emotional state and give appropriate feedbacks.

It is obvious that detecting the speaker's features is the foundation of the system. Performance of the speaker's emotional state detection will greatly influence the performance of our system. We propose to combine the speaker's facial expressions and speeches detected by camera and microphone to analyze his emotion. Although emotional facial expression and emotional speech recognition have developed for years, recognition results are not perfect for various emotions especially in real-time systems. We concern the implementation of our system, so we only divide emotional states into two kinds which are positive affection and negative affection and build two corresponding emotional backchannel lexicons. These two emotional states can be recognized correctly in current real-time experiments. Therefore we can apply the recognition technology in the backchannel system.

4 CONCLUSIONS

In face-to-face interaction between a human speaker and a virtual agent listener, addressing backchannel feedbacks to virtual agent listener gives the agent human-like conversation skills and creates rapport in Human-Computer Interaction. In recent years, researchers have made great efforts in predicting backchannel for the agent listener. In this position paper, we have argued the limitations of current approaches. It is time for us to look for new methods to improve the backchannel prediction and generation.

Following the two hypotheses emphasizing on the timing and type of backchannel, we introduce an improved system to enhance the agent listener's performance. In the backchannel prediction part, using Newcastle Personality Assessor before parasocial consensus sampling and neural networks will enable us to obtain the personality rules related to the number of backchannel. Then we can easily select different backchannel timing thresholds for specific agent listener's personality. In the backchannel generation part, we intend to build two emotional backchannel lexicons showing positive affection and negative affection respectively after conducting a context-free perceptual study. The system will randomly select one type of backchannel from the similar emotional backchannel lexicon in accordance with the empathy strategy. Further steps may include asking some volunteers to assess the system and keeping on developing it according to the evaluation results. These efforts will help to make the system more suitable for different conversation occasions. Implementation of the proposed system will greatly increase the naturalness between the human speaker and the agent listener.

ACKNOWLEDGEMENTS

This work is supported by the National Nature Science Foundation of China (No.61103097, No.60873269), International Science and Technology Cooperation Program of China (No.2010DFA11990).

REFERENCES

Bavelas, J. B., Coates, L., Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3), 566-580.

Bavelas, J. B., Coates, L., Johnson, T. (2000). Listeners as conarrators. *Journal of Personality and Social Psychology*, 79(6), 941-952.

Cathcart, N., Carletta, J., Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. *Proceedings of the Conference of the European chapter of the Association for Computational Linguistics*, 51-58.

Duncan, Starkey, Jr. and Fiske D. (1977). *Face-to-face Interaction.* New York: Halsted Press.

Duncan, Starkey, Jr. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(2): 161-180.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., Werf, R. J. V. D., Morency, L. (2006). Virtual Rapport. *Proceedings of the International Conference in Intelligent Virtual Agent*, 14-27.

Huang L., Morency, L., Gratch, J. (2010a). Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behaviour. *Proceedings of AAMAS 2010*, 1265-1272.

Huang L., Morency, L., Gratch, J. (2010b). Learning backchannel prediction model from parasocial consensus sampling. *Proceedings of the International Conference in Intelligent Virtual Agent*, 159-172.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26(1), 22-63.

Maatman, M., Gratch, J., Marsella, S. (2005). Natural behaviour of a listening agent. *Proceedings of the International Conference in Intelligent Virtual Agent*, 25-36.

McRorie, M., Sneddon, I., Sevin, E. D., Bevacqua, E., and Pelachaud, C. (2009). A Model of Personality and Emotional Traits. *Proceedings of the International Conference in Intelligent Virtual Agent*, 27-33.

Morency, L., Kok, I. D., Gratch, J. (2008). Predicting Listener Backchannels: A Probabilistic Multimodal Approach. *Proceedings of the International Conference in Intelligent Virtual Agent*, 176-190.

Poppe, R., Truong, K. P., Reidsma, D., et al. (2010). Backchannel strategies for artificial listeners. *Proceedings of the International Conference in Intelligent Virtual Agent*, 146-158.

Prendinger, H., Mori, J., Ishizuka, M. (2005). Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International Journal of Human Computer Studies*, 62(2), 231-245.

Ward, N., Tsukahara, W. (2000). Prosodic features which cue backchannel responses in English and Japanese. *Journal of Pragmatics*, 32(8), 1177-1207.

Yngve, V. (1970). On getting a word in edgewise. *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, 567-577.