Integrating Backchannel Prediction Models into Embodied Conversational Agents

Iwan de Kok and Dirk Heylen

Human Media Interaction, University of Twente
{i.a.dekok,heylen}@utwente.nl

Abstract. In this paper we will present our design for generating listening behavior for embodied conversational agents. It uses a corpus based prediction model to predict the timing of backchannels. The design of the system iterates on a previous design (Huang et al. [5]) on which we propose improvements in terms of robustness and personalization. For robustness we propose a variable threshold determined at run-time to regulate the amount of backchannels being produced by the system. For personalization we propose a character specification interface where the typical type of head nods to be displayed by the agent can be specified and ways to generate slight variations during runtime.

1 Introduction

One of the greatest challenges in developing embodied conversational agents is managing the flow of conversation. Succesful interaction between humans is achieved by complex coordination between verbal and nonverbal behaviors which together shape the information which is passed on from one interlocutor to the other. The behaviors that need to be displayed depend on the state the conversation is in.

With regards to turn-taking the conversational state of the agent will be modelled among two dimensions. The agent can *have* the turn or not and the agent can *want* the turn or not. These dimensions create four conversational states the agent can be in. Each of these states comes with their own type of actions and behaviors that are appropriate.

When the agent has and wants the turn, the agent will communicate what it has to share with its interlocutors. It will do this until it has shared all his information or until its turn is challenged by an interlocutor. At this time the agent will signal this through turn yielding behaviors. When it has lost the turn, the agent will need to display appropriate listening behavior to signal attendance to and understanding of its interlocutor. As soon as it wants the turn back it will need to display turn claiming behavior.

The design proposed in this paper is for the behavior of an agent in the conversational state that it does not have and does not want the turn. In this state the goal of our agent is to keep the interlocutor motivated in speaking by signalling attendance, understanding and/or appraisal through backchannels.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 268–274, 2012.

Humans succeeding in doing so increase quality of the speaker's speech [10,1], understanding of the speaker's speech by the listener [10,1] and rapport between the interlocutors [4].

This listening behavior is typically a combination of reactive and deliberative behavior. In our study we focus on reactive behaviors. Humans do not consciously plan each listener response, but they occur naturally without much thought. Over the past years several reactive prediction models have been developed to determine the timing of these listener responses. At first the handcrafted rules approach was utilized [16], but nowadays the corpus based machine learning approach has proven to outperform these handcrafted rules [12]. However, most implemented agents and robots still use these handcrafted rules [11,9,2] to time their listening behavior, the exception being [5].

Huang et al. [5] use a Conditional Random Fields model to predict the timing of the backchannels which is learned from a corpus of speaker-listener interactions. The timing of the backchannels is determed by comparing the output of the model to a threshold. The threshold is determined in the development of the model and optimized to optimally reproduce the behavior as observed in the corpus. At the predicted timings they randomly place one of three typical head nods, which were found in their corpus.

The proposed system in this paper intends to improve on this system in terms of robustness and personalization. For robustness we target the way the threshold is determined and used in the system. Due to changing conditions in an interactive system which can influence the output of prediction models, such as audio quality, recognition results of features or different speaking styles of users, a fixed threshold can lead to variable results. For some users the model will predict many backchannels, while for others it will predict hardly any backchannels. We propose a variable threshold determined at run-time to regulate the backchannel rate.

For personalization the proposed system allows for more types of head nods to be produced. It offers an interface to specify a character's typical head nods to define and personalize the created characters. Furthermore, these head nods are used as blue prints on which variations are generated. Depending on the certainty of the model that the prediction is correct, the system will generate a determined head nod on a high certainty prediction and a more shallow backchannel on a low certainty prediction.

In the remainder of the paper a general overview of the proposed system is given and the different components that are introduced there are discussed in more detail.

2 General Overview

Our design (see Figure 1) consist of two main components, the prediction module and the listener response generator. The prediction module monitors the interlocutor through the multimodal input channels of the system. Based on these observations, the model - which is trained on human-human interactions - produces

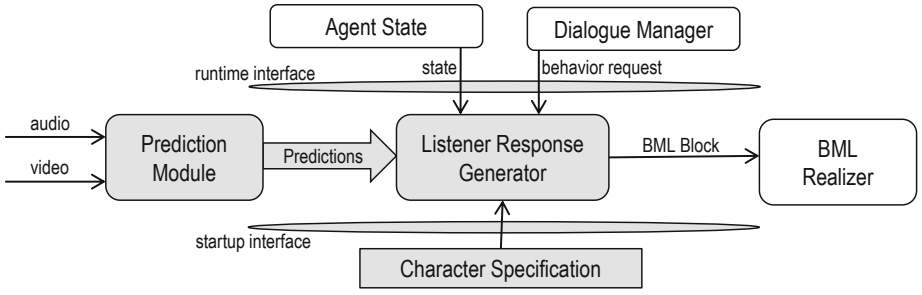

Fig. 1. Overview of the system architecture for generating listener responses. The gray parts and thick arrows are part of the system, while the thin arrows and white parts are part of the outside world, or part of the complete agent architecture.

a prediction value at each time frame, which indicates the appropriateness of a listener response at this time (see Section 4 for more details).

The listener response generator interprets these prediction values and decides at what times and with which form the listener responses are given. It generates BML blocks [15] which are passed on to a BML realizer (see Section 5).

There are two interfaces that can be used to influence the behavior of the system, one at startup and one during runtime. Through the interface at the startup a character's typical behavior can be specified (see Section 3.1). At runtime the agent's state is monitored which influences the behavior and the dialogue manager can request specific behavior from the system (see Section 3.2).

3 Interfacing with the Module

There are two ways to influence the behavior of the module. At the start up of the module the character specification is loaded. At runtime the module monitors the relevant states of the agent and it allows requests for specific behaviors.

3.1 Startup Interface

In our previous work we have seen that listeners differ greatly in amount of responses given, even when interacting in the same context as others [6]. Thus, the rate at which a listener responds is not only determined by the amount of opportunities given by the speaker and the understanding of the listener at these opportunities, but also by the individual characteristics of the listener.

Furthermore, it has been observed that listeners differ in the form they usually use as their listener response. Some listeners frequently use a vocal component in their responses, while others remain silent. Some listeners usually start their head nods upwards, others downwards and also the speed, amplitude and amount of nods per listener response differs between listeners.

To be able to easily generate listeners differing in these aspects with the same listening behavior module, a character specification is loaded at the startup of the module. This character specification constitutes the blueprint for the generated behavior of the listeners. In this specification the frequency of listener responses and the form of typical listener responses for this person are specified.

3.2 Runtime Interface

To allow the agent some control over the generated listening behavior by the listening behavior module a interface at runtime is available.

Through this interface the module monitors the relevant states of the agent. For the first implementation these states are limited to the conversational state (does the agent still neither have nor want the turn?) and the variable understanding. The conversational state determines whether the module produces behavior or not. Understanding has values between 0 and 1, where 0 is no understanding and 1 is full understanding. This state influences the form of the listener responses. Full understanding will produce determined listener responses and no understanding will display misunderstanding behavior.

The second way the module allows interaction is through the request channel. Through this channel requests can be made by the dialogue manager of the agent for a specific listener response. The module will generate this listener response at the first opportunity detected by the prediction module. The request has one of two priority labels; *high* or *low*. In the case of high priority the threshold the prediction value needs to exceed is lowered, while no manipulation of the threshold is performed in the case of low priority. This functionality will be implemented to allow more control from the agent over the generated behavior.

4 Listener Response Prediction Module

The listener response prediction model is responsible for predicting the timing of the generated listener responses. Contrary to the rule based prediction models used so far in embodied conversational agents, our listener response prediction model is a model trained on human-human conversations. Over the years many such models using various machine learning techniques, such as HMM [3], CRF [12,7,13] and SVM [8], have been trained and evaluated and proven to be more accurate than their handcrafted peers. Based on the input features describing the context these models make a prediction on how likely a listener response is at each moment in time. Input features typically used are eye gaze, prosody and lexical features and are derived from audio and video input. All these features can be detected and/or interpreted in real-time and incrementally.

For our implementation we will select an SVM model learned on the MultiLis corpus [8]. The implementation will be such that it is model independent and can easily be replaced by another model, as long as the output is a continuous stream of prediction values, similar to the output depicted in Figure 2.

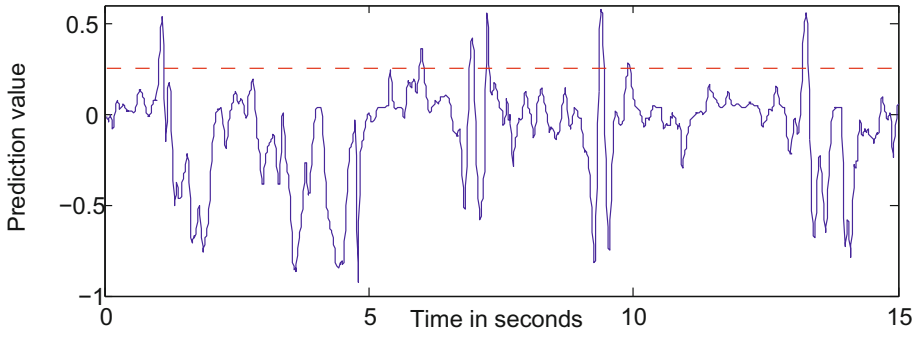

Fig. 2. The output of the prediction module; a prediction value curve

5 Listener Response Generator

The listener response generator is responsible for interpreting the output of the listener response prediction model and generating the appropriate response. The design will enable the module to generate varied, but personalizable listening behavior.

5.1 Timing of Listener Responses

The output of a listener response prediction model is a prediction value indicating the likelihood of a listener response occuring at each time frame. After sequencing and smoothing these prediction values one gets a prediction value curve as depicted in Figure 2. From this curve the timing of listener responses can be derived.

At certain times the prediction value curve peaks after a fast increase in prediction value. When the top of this peak exceeds a certain threshold (e.g. the red interrupted line) a listener response is predicted by the model. Selection of the threshold the peaks need to exceed is influenced by the response rate set in the character specification.

To create a more robust system with regards to the amount of backchannels that are generated, we will not select a fixed threshold like Huang et al. [5], but the threshold will be subject to change during an interaction. To regulate the response rate the threshold will slowly decrease as time since the last listener response goes by and increase as soon as a listener response is given by the agent. This will ensure that the system will be able to generate a similar amount of responses under changing conditions and for different speaking styles of users. Furthermore, this will ensure that long periods with no responses are less likely and that listener responses are not performed shortly after each other. This has been found to be perceived as erratic behavior for an agent [14], even though humans do this occasionally. The exact rates and type (linearly/exponentionally/...) of increase/decrease will be determined in the prototyping phase of development.

5.2 Form Selection

When a peak is detected exceeding the threshold, an appropriate behavior needs to be selected. If a request is made for a specific listener response, that listener response is generated (see Section 3.2). If no request is made for a specific listener response, the typical response for the character as specified by its specification (see Section 3.1) is generated.

Even though each individual has a preferred way of giving a listener response, these are not exactly the same. In fact, they vary their typical response slightly each time and at certain times they deviate from their typical response to give a really determined response.

We use the height of the peak in the prediction value curve to generate these variations in the form of the listener response. The higher the peak, the more determined the generated listener response is. This is achieved by increasing the amplitude and/or speed of the movement and/or the intensity of the facial expression. Since wrongly timed vocal listener responses are more often perceived as inappropriate than head nods [14], vocalizations are only added when the peak is high.

6 Conclusion and Future Work

In this paper we have presented our design for generating listening behavior for embodied conversational agents. The systems uses a corpus based prediction model to predict the timing of backchannels. The design of the system iterates on a previous design (Huang et al. [5]) on which we propose improvements in terms of robustness and personalization. For robustness we proposed a variable threshold determined at run-time to regulate the amount of backchannels being produced by the system. For personalization we propose a character specification interface where the typical type of head nods to be displayed by the agent can be specified and ways to generate slight variations during runtime.

For future work we intend to do a subjective evaluation of the system in which we evaluate the different components, by switching certain components and mechanics on or off.

References

1. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology 79(6), 941–952 (2000)
2. Bevacqua, E., McRorie, M., Pammi, S., Pelachaud, C., Schröder, M., Sneddon, I., de Sevin, E.: SAL multimodal generation component with customised SAL characters and visual mimicking behaviour. Tech. rep., SEMAINE Project (2009)
3. Fujie, S., Fukushima, K., Kobayashi, T.: A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In: Proc. Int. Conference on Autonomous Robots and Agents, pp. 379–384 (2004)

4. 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. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)
5. Huang, L., Morency, L.-P., Gratch, J.: Virtual Rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011)
6. de Kok, I., Heylen, D.: The MultiLis Corpus – Dealing with Individual Differences in Nonverbal Listening Behavior. In: Esposito, A., Esposito, A.M., Martone, R., Müller, V.C., Scarpetta, G. (eds.) COST 2010. LNCS, vol. 6456, pp. 362–375. Springer, Heidelberg (2011)
7. de Kok, I., Ozkan, D., Heylen, D., Morency, L.-P.: Learning and Evaluating Response Prediction Models using Parallel Listener Consensus. In: Proceeding of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (2010)
8. de Kok, I., Poppe, R., Heylen, D.: Iterative Perceptual Learning for Social Behavior Synthesis. Tech. rep., Centre for Telematics and Information Technology University of Twente (2012)
9. 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)
10. Kraut, R.E., Lewis, S.H., Swezey, L.W.: Listener responsiveness and the coordination of conversation. Journal of Personality and Social Psychology 43(4), 718–731 (1982)
11. Maatman, R.M., Gratch, J., Marsella, S.: Natural Behavior of a Listening Agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 25–36. Springer, Heidelberg (2005)
12. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems 20(1), 70–84 (2011)
13. Ozkan, D., Sagae, K., Morency, L.: Latent Mixture of Discriminative Experts for Multimodal Prediction Modeling. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 860–868. Association for Computational Linguistics (2010)
14. Poppe, R., Truong, K.P., Heylen, D.: Backchannels: Quantity, Type and Timing Matters. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 228–239. Springer, Heidelberg (2011)
15. Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: 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. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32(8), 1177–1207 (2000)