Controlling the Listener Response Rate of Virtual Agents

Iwan de Kok and Dirk Heylen

Human Media Interaction Group, University of Twente
P.O. Box 217, 7500AE Enschede, The Netherlands
{i.a.dekok,heylen}@utwente.nl

Abstract. This paper presents a novel way of interpreting the prediction value curves that are the output of the current state-of-the-art models in predicting generic listener responses for embodied conversational agents. Based on the time since the last generated listener response, the proposed dynamic thresholding approach varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. The proposed formula for this dynamic threshold includes a parameter which controls the response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to adapt the behavior to the situation, targeted role or personality of the virtual agent. We show that the generated behavior is more stable under changing conditions than the behavior of the traditional fixed threshold.

1 Introduction

In conversation, both speakers and listeners are active participants. Having a conversation requires complex coordination between verbal and nonverbal behavior to shape the information which is passed on from one interlocutor to the other. This is true for both the interlocutor who is speaking as well as the interlocutor currently listening. The speaker provides the information, while the listener is constantly providing feedback to the speaker by signallying his/her attendance, understanding and/or appraisal. This behavior is an essential part of a succesfull interaction. It has been proven to increase the quality of the speaker's speech [11,1], understanding of the speaker's speech by the listener [11,1] and rapport between the interlocutors [4].

The responses that interlocutors give while listening can be divided into generic and specific listener responses [1] (or alternatively continuers and assessments [3]). Generic listener responses are not specifically connected to what the speaker is saying. One could theoretically interchange two generic listener responses, typically head nods or minimal vocalizations such as "mhm", without having too much of an impact on the flow and meaning of the conversation. They merely function to signal attendence and a general notion of understanding to let the speaker know he/she can continue. Specific listener responses do have a tie to the content of the speaker's speech. They usually give an assessment

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 168–179, 2013.

of what has been said, such as "oh wow!" or gasping in horror, and/or give a specific signal of understanding, such as repeating key words of the speech.

Computational models of listening behavior for embodied conversational agents have also been developed and have shown success in replicating the function they fullfill in everyday conversation [12,10,19,6,18]. For the development of computational models for these and other embodied conversational agents different strategies for both types of listener responses have been considered. For specific listener responses the models have focused on the incremental understanding of the content of the speech [23,22], since these listener responses have strong ties to what is being said. Models for generic listener responses have focussed on more shallow features of the speaker's speech, such as acoustic features [2,24] and eye gaze [13]. The approach that has been utilized was initially handcrafted rules [24], but nowadays the corpus based machine learning approach has proven to outperform these handcrafted rules [13].

These machine learning based models are optimized to match the ground truth labels found in the corpus as closely as possible, usually measured by a F_1-measure [8]. The goal is generally not to copy the behavior that was found in a corpus, but to build an agent that responds to new input. Also, one might want to vary the behaviour of agent. So, when using such models in a conversational embodied agent, other aspects of the generated listening behavior may be more important. When designing a conversational embodied agent the designers usually have a personality or role in mind they want their agent to fulfill. A lot of factors are important when managing the impression a user has about the personality of an agent, one of which is their listening behavior. The timing, amount and form of the listener responses that are produced by an embodied conversational agent have been proven to influence the impression the user has about their personality [20].

Therefore, it is important that the produced listening behavior is consistent with the targeted personality of your embodied conversational agent and is so under every circumstance and for every user. In other words, it is important that the listening behavior that an embodied conversational agent performs is stable, recognizable and conform the expectations the user and designer have of the behavior.

This is typically not provided by the current state-of-the-art models for generating generic listener responses. Changes in conditions, such as different interlocutors with different voice characteristics and speaking styles, can have a big impact on the features used as input by these models, which in turn can have a big impact on the predictions made by the models.

In this paper we will present a novel way of interpreting the prediction value curves that are the output of the current state-of-the-art models for predicting generic listener responses for embodied conversational agents. Based on the time since the last listener response the proposed dynamic thresholding approach varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. The proposed formula for this dynamic threshold includes a parameter which controls the

response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to create the behavior that is desired for the targeted role, personality or situation.

More details about the dynamic thresholding approach are given in Section 2. We evaluate the approach on a corpus in Section 3. We conclude this paper with our final thoughts in Section 4.

2 Interpreting the Prediction Value Curve

The best known models to determine the timing of generic listener responses are corpus-based machine learning models [15,21,14,13,5,16]. These models are learned from an annotated corpus of human-human interactions. From this corpus features are extracted, such as the eye gaze of the speaker and speech features. Based on these features a model is learned that infers the relation between these features and the occurrence of a listener response in the corpus.

The output of such 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. In Figure 1 two example prediction value curves are presented, plotted in black (disregard the red line for now). These examples were taken from the first minute of two interactions from the MultiLis corpus and produced by the model we use in our evaluation. More details on the corpus and model follow later.

From these prediction value curves the timing predictions for listener responses can be extracted. This is usually done by detecting peaks in the prediction value curve and comparing these peaks to a threshold. If this peak exceeds the threshold, it is considered appropriate to give a listener response at the time of the peak. Typically the threshold is a fixed value that is determined in the validation phase of the development of the model, which can be decreased or increased to generate more or fewer responses respectively to express more attention or a different personality type. In Figure 1 the threshold that was found to give the highest F_1 score for the model is indicated by the horizontal line at 0.2122.

2.1 Limitations of the Fixed Threshold

However, this way of selecting the threshold has a problem. The amount of listener responses predicted by the model using a certain threshold is inconsistent. The same model using the same threshold applied to two different speakers can result in a significant variation in listener response rate which might be unwanted. This is illustrated by the two prediction value curves in Figure 1. By looking at the number of peaks that exceed the threshold we can see that applying the optimal threshold according to the validation step has resulted in big difference in response rate. For the first interaction nineteen listener responses are predicted in the first minute, while only one is predicted for the second interaction. The explanation for the lower prediction values in the bottom curve

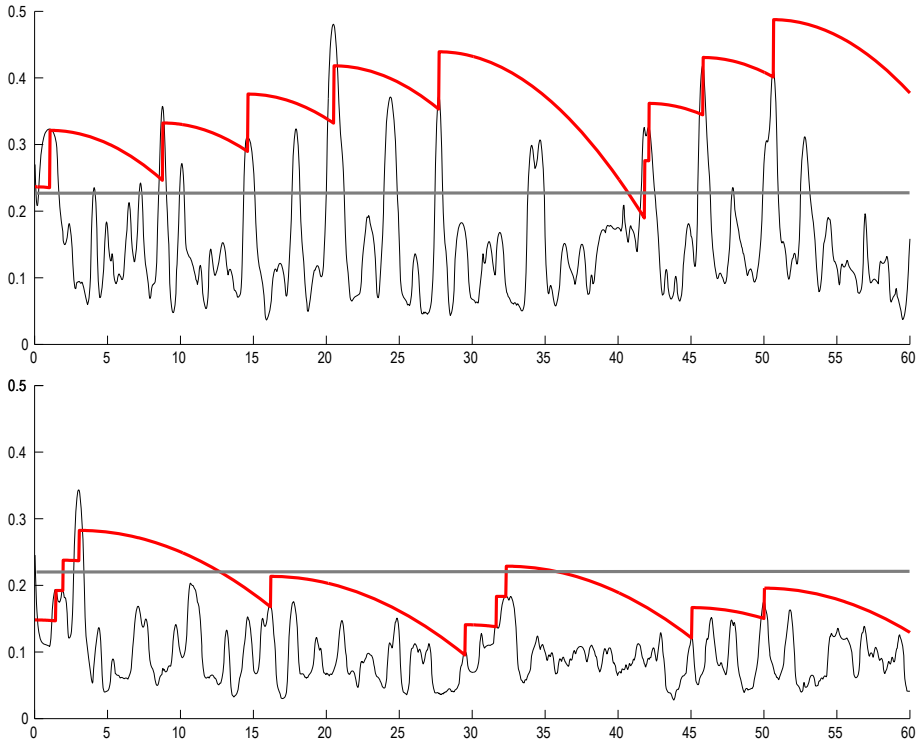

Fig. 1. The prediction value curves of the parallel listener consensus model applied to the first minute of two interactions. On the horizontal axis time is presented in seconds and on the vertical axis the likelihood of a listener response according to the model is presented. The gray horizontal line is the validated threshold obtained during the validation step. In red the proposed dynamic threshold is shown. The threshold start initially high and decreases over time at an increasing rate until a peak exceeds the threshold. Then the threshold is increased by a fixed amount after which the threshold start decreasing slowly again.

in Figure 1 lies in the fact that in this case the speaker hardly ever looked at the listener where gaze is one of the most important cues that the prediction model is using. Even though the speaker does not look often at the listener, opportunities to give a listener response are available, since the listeners in the corpus did respond during this minute. So, it might be in the interest of the virtual listener to give them. Either to comfort and encourage the speaker to continue speaking and/or to built a better rapport with the speaker.

With a fixed threshold this is hard to do, since their is no reliable way of knowing how much one needs to lower the threshold to get the desired response rate. This is because selecting peaks based on a fixed validated threshold is subject to changing conditions. These changing conditions do not limit themselves to eye gaze behavior as in the previous example, but also other aspects of a speaking style can change. Examples include, but are not limited to speaking

in a louder or softer voice, with higher or lower pitch, different speech rates or varying degrees in which intonation is used. These aspects can even change during an interaction with the same speaker, not necessarily because the speakers changes these characteristics of their speaking style, but they may also change their position relative to the microphone or video camera. All these things can influence the features used by the model, which in turn influence the prediction values returned by the model. For some conditions the general prediction value from a model will always be lower than for other conditions and the peaks in this curve may remain below the fixed validated threshold.

Essentially, these differing conditions makes it very difficult for a designer of an embodied conversational agent to give the agent the personality and behavior the designer has in mind. For one user the agent may be responsive and attentive, while for another almost no listener responses are generated at all.

So, we need another way of extracting the appropriate timings of listener responses from the prediction value curves. To give the designer of the nonverbal listening behavior of a virtual agent the tools to control the generated behavior, the solution needs to ensure the following characteristics:

- **Stable Response Rate** - The solution should be able to generate s similar response rate under changing conditions, such as different speakers, audio/video quality or feature extraction accuracy.
- **Evenly Distributed Responses** - Not only the overall rate should be stable, but the distribution needs to be even as well. Periods with many or few responses are perceived as unnatural behavior [17].
- **Adjustable Response Rate** - The response rate needs to be easily adjustable, so the designer of the agent can generate the desired behavior and even change this behavior during the interaction.

Next we will present a formula that is designed to adjust the thresholds dynamically according to these principles and we will show that it works.

2.2 Dynamic Thresholding

The solution we propose is to have a changing threshold during the interaction. So, instead of a fixed threshold determined at the development stage of the prediction model, we propose a dynamic threshold that changes over time depending on the time since the last predicted listener response. At the start of the listening period the threshold is relatively high and it starts decreasing at an increasing rate until a peak in the curve exceeds the threshold. When a listener response is predicted the threshold jumps up and starts decreasing again at an initially slow rate. This will ensure both the stable response rate and the even distribution of the responses.

In Figure 1 the dynamic threshold for the two interactions is shown in red. Here we can see that for both the example interactions the number of predicted response opportunities is nine. So, the resulting response rate is much more stable than the response rate the fixed threshold would have given, meaning we

have met our first required characteristic. The response are also more evenly distributed. There are no long periods without responses, the longest gap being 14 seconds. There are still issues with two or three consecutive predicted response opportunities, but an extra rule stating no two listener responses to be generated within a certain amount of time would solve this. For the final required characteristic we will take a closer look at the formula that created this dynamic threshold.

$$T_t = T_{t_{last}} + j - (\frac{t - t_{last}}{g.r})^2.\frac{j}{d} \quad (1)$$

The formula that created the dynamic thresholds presented in red in Figure 1 is presented in Equation 1. It calculates the dynamic threshold (T) at time t. Time t is measured in frames. Gap parameter g is the mean time in seconds between two predicted listener responses. Parameter r is the sampling rate of the system, needed to convert timing in seconds into timing in frames. Parameter t_{last} is the time of the listener response that was last generated and $T_{t_{last}}$ is the dynamic threshold at that time. Parameter j is the jump parameter, which represents the amount that the dynamic threshold increases after predicting a listener responses. The final parameter d is the dropoff parameter, which controls the amount the dynamic thresholds decreases over time.

Gap parameter g is the parameter that will give the designer of the agent control over the response rate of the agent by defining the mean gap between two predicted listener responses. However, before this gap parameter will give the expected behavior, the jump parameter j and dropoff parameter d need to be determined. For the jump parameter j we recommend using the standard deviation of the prediction value curve as value. This will make the jumps after predicting a listener response appropriate to the variation found in the prediction value curve and thus make the dynamic threshold even more adaptable to differing conditions. That leaves the dropoff parameter and this needs to be calibrated on a development set of example interactions. The procedure for this is to try different combinations of parameter d and g and minimize the absolute difference in expected number of listener responses and predicted number of listener responses for the values of g you expect the virtual listener to use.

3 Evaluation

In the previous section we have presented our dynamic threshold formula and highlighted its merits on two segments of one minute. In this section we will evaluate the performance of the dynamic threshold over a larger sample size. We will do this on a subset of the interactions of the MultiLis corpus, presented in Section 3.1, using a prediction model, presented in Section 3.2, that is trained on the other subset of interaction of the MultiLis corpus. The procedure of this evaluation is explained in Section 3.3.

To support the dynamic thresholding formula the evaluation will aim to give answers to the following questions: Does the dynamic threshold succeed in stabilizing the response rate? Does it avoid periods with few or many responses?

3.1 MultiLis Corpus

The MultiLis corpus [7] is a Dutch spoken multimodal corpus of 32 mediated face-to-face interactions totaling 131 minutes. Participants (29 male, 3 female, mean age 25) were assigned the role of either speaker or listener during an interaction. In each session four participants were invited to record four interactions. Each participant was once speaker and three times listener.

The interactions of the corpus were between one speaker and three listeners. The three listener were tricked into believing to be the sole listeners, but were recorded in parallel listening to the same speaker. The speakers saw only one of the listeners, believing that they had a one-on-one conversation. To create this illusion all listeners were placed in a cubicle and saw the speaker on the screen in front of them. The camera was placed behind an interrogation mirror, positioned directly behind the position on which the interlocutor was projected. This made it possible to create the feeling of eye contact.

To ensure that the illusion of a one-on-one conversation was not broken, interaction between participants was limited. Speakers and listeners were instructed not to ask for clarifications or to elicit explicit feedback from each other, so no turn-switching would take place. The speaker received a task of either watching a short video clip before the interaction and summarizing it to the listener, or learning a recipe in the 10 minutes before the interaction and reciting it to the listener. The listener needed to remember as many details of what the speaker told as possible, since questions about the content were asked afterwards.

In our evaluations we use ten interactions from the corpus totalling little over 40 minutes. These ten interactions were not used in the development of the prediction model, which will be presented in the following section.

3.2 Prediction Model

As our listener response prediction model we use the best performing model from the paper by De Kok et al. [9]. This model is the "Consensus 2" model. It is a Conditional Random Fields model trained on the other 22 interactions from the MultiLis corpus. As input features it uses the eye gaze of the speaker and several acoustic features of the speech signal, such as pitch, intensity and silence. As ground truth labels the model is trained using only the moments where at least two of the three listeners in the MultiLis corpus have given a listener response.

3.3 Procedure

We applied the listener response prediction model to the ten interactions to obtain the prediction value curve for each interaction. We then applied eleven fixed and eleven dynamic thresholds to these prediction value cur-ves. We varied the fixed thresholds between 0.15 and 0.35. To have dynamic thresholds we can directly compare to the fixed thresholds, we look at the resulting overall response rate of the thresholds. We then set gap parameter g from the dynamic threshold

Table 1. The table illustrates the effect of different fixed thresholds on the response rate in responses per minute of a listener response prediction model. The cells are gray shaded for easier interpretation, with higher response rate being darker. It illustrates that, although increasing the threshold decreases the overall response rate at a predictable pace, the effects on individual interactions varies wildly.

	Response Rate for Fixed Threshold (responses/minute)										
		Interaction									
Threshold	**Overall**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.15	21.6	19.0	23.8	26.1	21.9	12.9	21.8	24.9	24.2	22.7	21.5
0.17	16.8	15.7	20.2	21.6	17.2	6.4	16.7	22.3	19.2	16.4	16.1
0.19	13.2	13.7	16.2	18.6	12.1	3.0	12.0	18.4	15.5	13.4	13.1
0.21	11.5	12.8	14.8	16.7	11.1	2.0	8.8	15.3	13.0	12.5	11.5
0.23	10.1	11.3	13.3	15.7	10.0	1.5	7.2	14.4	11.3	10.8	10.9
0.25	8.7	9.3	11.5	14.1	8.7	1.0	5.5	13.6	9.8	9.6	8.4
0.27	7.1	7.0	10.4	13.4	8.0	1.0	3.5	11.8	8.2	7.2	6.6
0.29	5.5	4.0	8.6	12.4	6.7	1.0	2.9	10.9	6.3	5.3	4.1
0.31	4.4	3.1	6.1	10.1	5.4	0.7	2.1	10.1	5.3	3.8	3.2
0.33	3.1	2.2	4.3	5.9	4.1	0.7	1.9	8.7	2.8	2.8	2.7
0.35	2.4	1.8	2.5	4.9	3.1	0.5	1.5	6.6	2.3	1.3	2.5

formula such that the response rate of that threshold was (almost) the same. For both the fixed and the dynamic threshold we do not allow any listener response within one second of the previous listener response. These predicted listener responses are discarded.

For the dynamic threshold other parameter beside the gap parameter g need to be set as well. For this evaluation we initialized the dynamic threshold with the mean of the prediction value curve as initial $T_{t_{last}}$ and the standard deviation of the prediction value curve as jump parameter j. To select the value for the dropoff parameter d, we tried several combinations of gap parameter g and dropoff paramater d on the ten interactions. We selected the value for dropoff parameter d such that the difference between the expected number of listener responses based on the value for gap parameter g and the resulting number of predicted listener responses was minimized. This was true for value 1.4.

3.4 Results

The first question we will answer is, does the dynamic threshold succeed in stabilizing the response rate? For this we look at the response rates that were the result of applying the eleven fixed and dynamic thresholds on each interaction. These response rates in responses per minute are presented in Tables 1 (fixed thresholds) and Table 2 (dynamic thresholds). In the first column the height of the fixed threshold and the gap parameter g that were varied are presented. A comparison of the second columns of both Table 1 and Table 2 shows that a similar overall response rate is predicted by the model by the both thresholds.

In the remaining columns the response rates for each interactions are presented. The cells are gray shaded for easier interpretation, with higher response

Table 2. The table illustrates the effect of different gap parameters in the dynamic threshold on the response rate in responses per minute of a listener response prediction model. The cells are gray shaded for easier interpretation, with higher response rate being darker. It illustrates that, for each interaction a similar response rate is generated.

	Response Rate for Dynamic Threshold (responses/minute)										
		Interaction									
Gap g	**Overall**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2.8	18.1	17.7	19.1	18.6	18.0	17.8	18.4	19.2	17.0	18.5	18.1
3.6	14.8	14.1	14.8	15.4	14.7	14.4	14.6	15.7	14.5	15.1	15.2
4.5	12.1	11.9	12.2	12.4	12.1	11.9	11.9	13.1	12.0	12.1	12.2
5.2	10.6	10.4	10.4	10.8	10.5	10.2	11.0	11.4	10.5	10.8	10.6
5.9	9.5	9.8	9.4	10.1	9.3	8.9	9.3	10.1	9.2	9.8	9.9
6.9	8.4	8.8	8.3	8.5	8.0	7.9	8.3	9.2	8.2	8.7	8.6
8.4	7.0	6.8	7.6	7.5	6.9	6.7	6.7	7.9	6.7	7.4	7.2
10.9	5.7	5.6	5.8	6.2	5.9	5.2	5.5	6.6	5.3	5.7	5.7
13.7	4.7	4.3	4.3	5.2	4.9	4.7	4.5	5.2	4.5	4.9	4.7
19.2	3.6	3.3	4.0	3.6	3.6	3.5	3.4	3.9	3.3	3.8	3.6
25.2	2.9	2.6	3.2	3.1	3.1	3.0	2.8	3.5	2.7	3.2	2.9

rate being darker. These rates show that for the fixed threshold the response rate of each interaction can vary wildly. Especially the predicted response rate for interaction 5 is a lot lower than for the other interactions, while the predicted response rate for interaction 7 is generally higher. For the dynamic threshold the response rates for each individual interaction are a lot more stable. Each interaction has more or less the same response rate. So indeed, the dynamic threshold has succeeded in stabilizing the response rate.

The next question we will answer is, does the dynamic threshold avoid periods with many or few responses? First we look at whether the dynamic threshold avoids periods with many responses. For this, we will see how many responses were discarded by the extra rule that no predicted responses should be within 1 second of each other. The number of discarded predictions is presented in Table 3 and the left graph in Figure 2. They show that this number is about twice as high in case of the fixed threshold compared to the dynamic threshold. So, the dynamic threshold helps to avoid periods with many responses.

To see whether the dynamic threshold helps to avoid periods with few predicted responses, an analysis of the gaps between two consecutive predicted responses is performed. We looked at the mean, standard deviation and maximum of these gaps. The results of this analysis are presented in Table 3. By looking and comparing the columns μ and max of both thresholds, we can see that the difference between the mean and maximum gap is a lot more stable for the dynamic threshold (see also right graph in Figure 2). This is also reflected in the lower standard deviation for the dynamic threshold (see middle graph in Figure 2). So, we can conclude that the dynamic threshold also resulted in a more even distribution of the predicted responses.

Table 3. Presentation of the results of the analysis of the gaps between predicted responses using a fixed threshold and our dynamic threshold

Fixed Threshold					Dynamic Threshold				
	Discarded	Gap (s)				Discarded	Gap (s)		
Threshold	Predictions	μ	σ	max	Gap g	Predictions	μ	σ	max
0.15	384	2.75	1.85	24.17	**2.8**	129	3.31	1.41	7.33
0.17	202	3.56	2.93	47.10	**3.6**	75	4.06	1.89	9.37
0.19	114	4.51	5.12	85.63	**4.5**	52	4.97	2.53	10.93
0.21	79	5.21	6.06	85.63	**5.2**	43	5.65	2.98	13.67
0.23	62	5.91	6.68	85.63	**5.9**	40	6.35	3.42	15.17
0.25	55	6.94	8.25	97.60	**6.9**	27	7.17	4.09	17.13
0.27	41	8.51	9.60	97.60	**8.4**	18	8.61	4.95	20.97
0.29	23	11.01	12.48	97.60	**10.9**	16	10.74	6.79	27.03
0.31	16	13.73	15.56	97.60	**13.7**	11	13.13	8.30	33.47
0.33	11	19.15	19.77	101.93	**19.2**	6	17.53	12.13	43.63
0.35	9	22.80	22.61	101.93	**25.2**	5	21.49	16.11	55.87

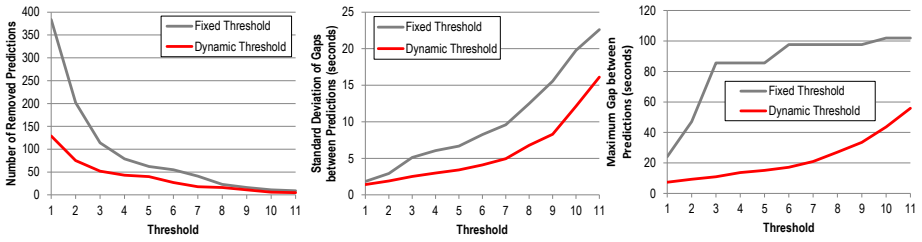

Fig. 2. Three graphs to illustrate that for all thresholds the dynamic thresholding needs less predictions to be removed (left graph), has a lower standard deviation for the gaps between two consecutive predictions (middle graph) and has a smaller maximum gap between consecutive predictions (right graph)

4 Conclusion

In this paper we have presented a novel way of interpreting the prediction value curves that are the output of the current state-of-the-art models in predicting generic listener responses for embodied conversational agents. Based on the time since the last generated listener response the dynamic thresholding approach varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. We have shown through an objective evaluation that this approach generates behavior that is more stable under changing conditions.

Furthermore, the proposed formula for this dynamic threshold includes a parameter which controls the response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to create the behavior that is desired for the targeted role, personality or situation. When the

role, personality or situation requires a low listener response rate, the designer now has a reliable way of ensuring this response rate for any interlocutor.

A subjective evaluation of the generated behavior has been performed, but full coverage of this evaluation was considered outside the scope of this paper. In this subjective evaluation the behavior generated with the dynamic thresholding was preferred over the behavior generated with the fixed threshold.

The current state-of-the-art models for predicting generic listener responses are general models aimed to work for every situation, for every interaction partner and for every other context one can think of. The dynamic threshold proposed in this paper is a way of achieving this with the current state-of-the-art prediciton models, but ultimately we need more advanced models that can adapt to different speaking styles, conversational settings and/or changing conditions by themselves.

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. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. European ACL pp. 51–58 (2003)
3. Goodwin, C.: Between and within: Alternative sequential treatments of continuers and assessments. Human Studies 9(2-3), 205–217 (1986)
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.: Learning Backchannel Prediction Model from Parasocial Consensus Sampling: A Subjective Evaluation. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 159–172 (2010)
6. Huang, L., Morency, L.P., Gratch, J.: Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In: Proceedings of Autonomous Agents and Multi-Agent Systems, Toronto, Canada, pp. 1265–1272 (2010)
7. 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 2102 Int. Training School 2010. LNCS, vol. 6456, pp. 362–375. Springer, Heidelberg (2011)
8. de Kok, I., Heylen, D.: A survey on evaluation metrics for backchannel prediction models. In: Interdisciplinary Workshop on Feedback Behaviors in Dialog, pp. 15–18 (2012)
9. 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)
10. Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T.: Modeling Embodied Feedback with Virtual Humans. In: Wachsmuth, I., Knoblich, G. (eds.) Modeling Communication. LNCS (LNAI), vol. 4930, pp. 18–37. Springer, Heidelberg (2008)

11. 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)
12. 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)
13. 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)
14. Nishimura, R., Kitaoka, N., Nakagawa, S.: A spoken dialog system for chat-like conversations considering response timing. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 599–606. Springer, Heidelberg (2007)
15. Noguchi, H., Den, Y.: Prosody-based detection of the context of backchannel responses. In: Fifth International Conference on Spoken Language Processing (1998)
16. Ozkan, D., Morency, L.P.: Latent Mixture of Discriminative Experts. IEEE Transaction on Multimedia 15(2), 326–338 (2013)
17. Poppe, R., Truong, K.P., Heylen, D.: Perceptual evaluation of backchannel strategies for artificial listeners. Autonomous Agents and Multi-Agent Systems (January 2013)
18. Sakai, Y., Nonaka, Y., Yasuda, K., Nakano, Y.I.: Listener agent for elderly people with dementia. In: Proceedings of HRI 2012, pp. 199–200 (2012)
19. Schröder, M., Bevacqua, E., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., Pammi, S., Pantic, M., Schuller, B., Pelachaud, C., de Sevin, E., Wollmer, M., Valstar, M.: A demonstration of audiovisual sensitive artificial listeners. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–2. IEEE, Amsterdam (September 2009)
20. de Sevin, E., Hyniewska, S.J., Pelachaud, C.: Influence of personality traits on backchannel selection. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 187–193. Springer, Heidelberg (2010)
21. Takeuchi, M., Kitaoka, N., Nakagawa, S.: Timing detection for realtime dialog systems using prosodic and linguistic information. In: International Conference on Speech Prosody, pp. 529–532 (2004)
22. Traum, D., DeVault, D., Lee, J., Wang, Z., Marsella, S.: Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 275–288. Springer, Heidelberg (2012)
23. Wang, Z., Lee, J., Marsella, S.: Towards More Comprehensive Listening Behavior: Beyond the Bobble Head. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 216–227. Springer, Heidelberg (2011)
24. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32(8), 1177–1207 (2000)