Investigating the influence of pause fillers for automatic backchannel prediction

Stefan Scherer[1], *Derya Ozkan*[1], *Louis-Philippe Morency*[1]

[1] Institute of Creative Technologies, University of Southern California, United States

scherer@ict.usc.edu

1. Introduction

Hesitations, and pause fillers (e.g. "um", "uh"), occur frequently in everyday conversations or monologues. They can be observed for a wide range of reasons including: lexical access, structuring of utterances, and requesting feedback from the listener [1]. In this study we investigate the usefulness of pause fillers as a feature for the prediction of backchannels using conditional random fields (CRF) [2] within a large corpus of interactions.

Backchannel feedbacks (i.e. the nods and paraverbals such as "uh-hu" and "mm-hmm" that listeners produce as someone is speaking) play a significant role in determining the nature of a social exchange by showing rapport and engagement [3]. When these signals are coordinated and reciprocated, they can lead to feelings of rapport and promote beneficial outcomes in diverse areas such as negotiations and conflict resolution [4], psychotherapeutic effectiveness [5], improved test performance in classrooms [6] and improved quality of child care [7]. Therefore, the prediction of backchannel feedback can play a significant role in a range of applications. For virtual human systems for example the correct timing of backchannels could be used to signal active listening or interest in the conversation with the human interlocutor. Additionally, one could provide systems with a stronger sense of rapport.

The remainder of the paper is organized as follows: in Section 2 we introduce the dataset utilized in the study. Section 2.1 statistically evaluates the relation between backchannel feedback and hesitations, revealing a rough sense of the applicability of hesitations for the prediction of backchannels. Section 3 reports the conducted experiments for backchannel prediction and reports the achieved results. Finally, Section 4 discusses the results and provides an outlook for further investigations.

2. Dataset

In this study we utilized a large dataset of 43 unique interactions[1]. The data was recorded in human-human interactions with two unique interlocutors in each conversation [8]. One participant was instructed to be the listener while the other person narrated a video clip taken from a sexual harassment awareness video by Edge Training Systems.

Synchronized multimodal data from each participant including voice and video were collected. Both the speaker and listener wore a lightweight headset with microphone. The average signal to noise ratio is very low at about 11.95 dB, indicating a relatively high level of noise within the data.

Human coders manually annotated the narratives, including pauses, hesitations, i.e. filled pauses (e.g. "um", "uh"), as well as incomplete and prolonged words; the transcriptions were double-checked by a second transcriber.

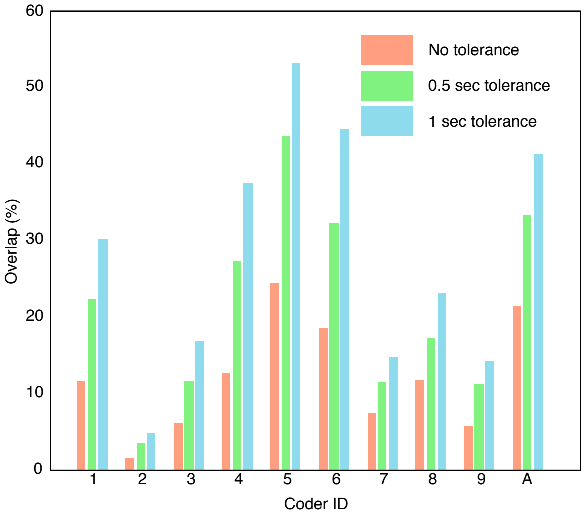

Figure 1: Overlap in percentage of hesitations and backchannel feedback for all nine coders (1-9) and actual feedback by original listeners (A).

In the present study we focused on the annotated hesitations and backchannels. The vocabulary of hesitations includes the following words: "um", "uh", "er" and "ah". In total we found 470 such annotations in the dataset with an average length of 0.32 seconds (0.12 standard deviation) uttered by 50 unique speakers. The rest of the words within the dataset have an average length of 0.29 seconds (0.17 standard deviation).

In total we observed 690 backchannels within the conversations, whereas the feedback behavior of each listener varied a lot. In order to provide the automatic prediction model with more homogenous training data, we employed the parasocial consensus sampling (PCS) paradigm [9], which enables efficient label acquisition from multiple coders. PCS is applied by having participants watch pre-recorded videos drawn from the RAPPORT dataset. In [9], nine participants were recruited, who were told to pretend they are an active listener and press the keyboard whenever they felt like providing backchannel feedback. This provides us with the responses from multiple listeners all interacting with the same speaker.

2.1. The statistical relation between hesitations and backchannels

In this section we investigate the statistical relations between backchannels and hesitations. As mentioned above we found 690 backchannels produced by the actual listeners and 319 hes-

[1] http://rapport.ict.usc.edu

itations uttered by the speakers in 43 unique interactions. The nine additional coders provided on average 644.7 backchannels. In Figure 1 the percentage of overlapping hesitations and backchannels are listed for all the coders and the actual backchannels. The percentage is calculated with respect to the total number of hesitations. Additionally, we show varying so called "tolerance"-levels. Level 0 means that the hesitation has to overlap with the backchannel, level 0.5 indicates that the hesitation can be delayed or preceding the backchannel by 0.5 seconds and level 1.0 respectively means that the hesitation can be delayed or preceding the backchannel by 1 second.

It is seen that several coders, as well as the actual backchannel timings overlap significantly with the hesitations. Coders 4 through 6 have high percentage numbers of overlap and improvements in the backchannel prediction experiments is suspected for those coders. Coders with very little overlap, such as Coder 2, probably do not take hesitations into account when providing backchannel feedback. Therefore, no improvement is to be expected for those coder's backchannel prediction.

3. Backchannel prediction experiments

The experiments are based on the CRF approach found in [2]. We combined multimodal features in one large feature vector for the CRF model along with the hesitation timings. To be precise, the utilized multimodal features included the following: Eye gaze, lowness (i.e. low pitch values), head nods, pause timings and smiles.

We performed hold-out testing on a randomly selected subset of ten interactions. The training set contains the remaining 33 interactions. Model parameters were validated by using a three-fold cross-validation strategy on the training set.

3.1. Experimental results

In the experiments, the CRF needs to decide for each input frame if a backchannel will follow or not. We evaluate the performance of the CRF using a slightly modified version of the F_1 measure, which is the weighted harmonic mean of precision and recall. Precision is the probability that predicted backchannels correspond to actual listener behavior. Recall is the probability that a backchannel produced by a listener in our test set was predicted by the model.

We first find all the "peaks" (i.e., local maxima) from output probabilities. If a peak coincides with an actual backchannel, then it is counted as a *hit*. If the peak is outside the boundaries of a backchannel it is counted as a *false alarm* and if no such instance is found within the borders of a backchannel it counts as a *miss*. We compare the performance of CRF utilizing the previously mentioned feature set with and without hesitations as an additional feature.

Figure 2, summarizes the performances of the experiments for the different coders. It is seen that for about half of the coders the performance improved, whereas for the other half the performance declined. It is worthy to note, that the best performing models i.e. coders 4 (F_1 without hesitations: 0.317; F_1 with hesitations: 0.321), 5 (F_1 without hesitations: 0.442; F_1 with hesitations: 0.449) and 6 (F_1 without hesitations: 0.352; F_1 with hesitations: 0.361) slightly improved the results and the hesitations therefore allowed for an improvement of the present baseline. Unfortunately, the results do not provide significant improvements.

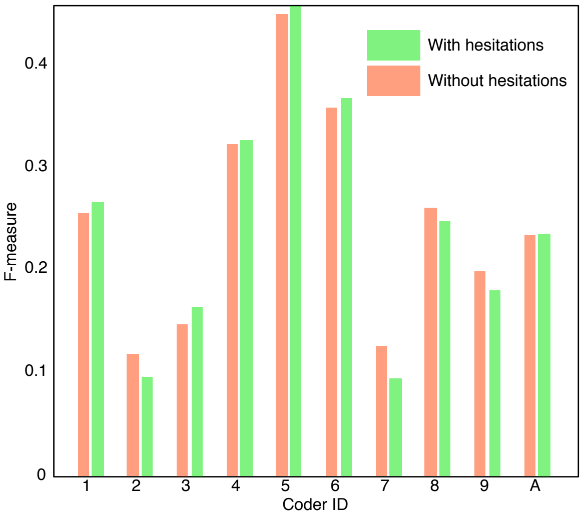

Figure 2: F_1 scores for the conditional random field backchannel feedback prediction with and without the additional feature of hesitation for the nine individual coders.

4. Summary

In this study we investigated the influence of hesitations for the automatic prediction of backchannels using CRF models. We compare the performance of the models with and without using hesitations as an additional feature for the prediction. We could see improvements for several coders, however, no clear trend could be found. We can confirm that different listeners utilize different cues for the decision when to provide a backchannel. The variations of listener behavior should be investigated further.

5. References

[1] R. Carlson, K. Gustafson, and E. Strangert, "Modelling hesitation for synthesis of spontaneous speech," *Proceedings of Speech Prosody, Dresden, Germany*, 2006.

[2] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in *Conference on Intelligent Virutal Agents (IVA)*, 2008.

[3] J. Gratch, N. Wang, J. Gerten, and E. Fast, "Creating rapport with virtual agents," in *Intelligent Virtual Agents (IVA)*, 2007.

[4] A. L. Drolet and M. W. Morris, "Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts," *Journal of Experimental Social Psychology*, vol. 36, no. 1, pp. 26–50, 2000.

[5] P. Tsui and G. Schultz, "Failure of rapport: Why psychotheraputic engagement fails in the treatment of asian clients," *American Journal of Orthopsychiatry*, vol. 55, pp. 561–569, 1985.

[6] D. Fuchs, "Examiner familiarity effects on test performance: implications for training and practice," *Topics in Early Childhood Special Education*, vol. 7, pp. 90–104, 1987.

[7] M. Burns, "Rapport and relationships: The basis of child care," *Journal of Child Care*, vol. 2, pp. 47–57, 1984.

[8] L.-P. Morency, I. Kok, and J. Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in *Proceedings of the 8th international conference on Intelligent Virtual Agents*, ser. IVA '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 176–190.

[9] L. Huang, L.-P. Morency, and J. Gratch:, "Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010.