INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Feedback Utterances

Catharine Oertel[1], *Joakim Gustafson*[1], *Alan W. Black*[2]

[1]KTH Royal Institute of Technology
[2] Carnegie Mellon University
catha@kth.se, jocke@speech.kth.se, awb@cs.cmu.edu

Abstract

Current speech synthesizers typically lack backchannel tokens. Those synthesiser, which include backchannels, typically only support a limited set of stereotypical functions. However, this does not mirror the subtleties of backchannels in spontaneous conversations. If we want to be able to build an artificial listener, that can display degrees of attentiveness, we need a speech synthesizer with more fine–grained control of the prosodic realisations of its backchannels.

In the current study we used a corpus of three-party face-to-face discussions to sample backchannels produced under varying conversational dynamics. We wanted to understand i) which prosodic cues are relevant for the perception of varying degrees of attentiveness ii) how much of a difference is necessary for people to perceive a difference in attentiveness iii) whether a preliminary classifier could be trained to distinguish between more and less attentive backchannel token.

Index Terms: conversational speech, backchannels, attentiveness

1. Introduction

Humans spend a substantial part of their day listening to other humans. Parents listen to their childrens' latest adventures, therapists listen to their patients, and professionals listen to each other during company briefings etc. While the contexts which surround the listening can vary a lot, the function still remains the same; an attentive listener lends support to the speaker and provides him with feedback.

Like with humans, the contexts under which an artificial listener could be useful, are manifold. However, especially for systems focusing on building artificial companions, a realisation of human-like feedback behaviour could be very useful. Further application areas could include companions specialised for interactions with the elderly [1, 2] or artificial listener to function as an interviewer for healthcare decision support [3].

In recent years more and more research has focused on investigating listening behaviours. Most of the research in this domain has focused on developing models to predict the correct timing of feedback utterances, as for instance described in [4, 5, 6].There are also some studies, which set out to describe the different functions, feedback tokens could take on, and how these functions are prosodically realised [7, 8].To our knowledge, there is however only one study, which focuses on describing differences in the realisation of feedback tokens between distracted and attentive listeners [9].

The aim of the current study is to lay the groundwork for a conversational speech synthesiser, which is able to increase or decrease the amount of perceived attentiveness, on the granularity of the smallest perceivable difference. For this we first carry out an in depth analysis of third-party observer agreement in relation to relative prosodic differences 5.1. Second, we investigate whether there are any significant difference between backchannel tokens which are rated to be more attentive in comparison to the backchannel tokens which were rated to be less attentive 5.2. Finally, in section 5.3 we propose a preliminary classifier which is focused on ranking bisyllabic backchannel, rather then classifying them according to predefined groups.

2. Background

There are only a few studies that investigates the prosodic realization of backchannel functions. Neiberg et al.[8] investigated how prosodic realization influence the perceived function of feedback tokens, taken from dyadic conversations. They found that feedback tokens were often multi-functional, some conveyed both understanding, agreement, certainty, and negative surprise. The perceived function was found to be correlated with prosodic cues; e.g. tokens with a fast speaking rate and a moderate F0 rise were found to convey understanding and interest. A further study on the prosodic characteristics of German feedback expressions was done by Malisz et al. [9]. They analysed the prosodic characteristics of ("ja", "m" and "mhm") across their pragmatic functions as well as the differences in feedback produced by distracted vs. attentive listeners. They split each feedback signal into three parts of equal length and calculated the mean and standard deviation for each of these parts. They calculated slopes over the first and the second half. They then used Generalised Linear Mixed Models (GLMM) for investigating feedback function differences, and distractedness-related differences. They found that attentive speakers tend to speak more loudly, energy is less variable, and pitch variability measures are positively related to attentiveness. They argue that prosodic features may strongly depend on segmental structure eg. nasality vs orality syllabic structure vs monosyllabic structure. Ward [7] found that syllabification, duration, loudness, pitch slope and pitch contour are relevant for the functional feedback categories in English. Kawahara et al. [10] try to predict an audience's interest level on the basis of backchannel tokens. For this they also analyse the prosodic realisation of japanese feedback tokens. They find that prolonged "hu:N" means interest and surprise while "a:" with higher pitch or larger power means interest. On the other hand, "he:" can be emphasized in all prosodic features to express interest and surprise. Gustafson and Neiberg [11] analysed feedback tokens over the course of a Swedish-call-in-radio-show. They found that interest-signalling, and encouraging pitch contours, are most commonly found at the beginning of the call. Over

http://dx.doi.org/10.21437/Interspeech.2016-1274

time the mean intensity of the feedbacks tokens decreases, and bisyllabic tokens become flatter, which basically turns them into monosyllabic feedback tokens. They also note that the overall pitch level decreases.

3. Data

3.1. Corpus

The KTH-Idiap corpus [12] is a corpus of multi-party group discussions. Three PhD students had to convince a Post-Doc that they were the best suited candidate for a prestigious scholarship. We designed the corpus to encompass four distinct phases. During the first phase, the students were asked to shortly introduce themselves. In the second phase, they were asked to give an elevator-pitch-about their respective PhD project. The fourth phase consisted of the students describing their projects' impact on society, and in the fifth and final phase they were asked to collaboratively come up with a joined project proposal. In a previous study, on the same corpus [13], it was found that third party observers could distinguish between at least three distinct types of listeners; an attentive listener, a side-participant, and a bystander. It was also found that the frequency of backchannel tokens was related to the perception of listener categories. No analysis however, was carried out on the prosodic realisation of these backchannel token. As the corpus is rich in conversational dynamics in general, and listener categories in particular, we chose to use it for this study.

3.2. Feature Extraction

In order to determine syllable boundaries, syllables were manually annotated in Praat [14]. We then calculated the syllable duration, in milliseconds, from the corresponding TextGrid file and, in a next step, extracted pitch and intensity values using Praat [14]. We converted Hz values into semitones and normalised F0-mean as well as rms-intensity values by speaker.

4. Perception Test Setup

4.1. Stimuli

The dataset we used in this study consists of bisyllabic as well as mono-syllabic backchannel token. We include 254 unique pairs of bisyllabic backchannel tokens, containing 64 unique backchannels. We also include 197 unique pairs of monosyllabic backchannel token, containing 91 unique backchannels. We sampled the backchannels across 9 speakers (5 male, 4 female). Each backchannel pair was rated by 24 raters.

To avoid cross-speaker confounds, pairs were only created of within speaker comparisons. Each item, of a pair of backchannel tokens, was embedded into the same carrier sentence, so that it was possible to ensure that backchannel tokens were rated in the same interactional environment. All carrier sentences, as well as all backchannel tokens, were from the KTH-IDIAP group conversation corpus. Backchannels were inserted into the same place in which a backchannel had occurred in the original recordings.

4.2. Raters

Raters were recruited from the crowdsourcing platform crowdflower. We restricted the geographic area of the rater to the United States, The Netherlands and Germany. No effect of geographic area on the ratings was found. To ensure that we received the highest possible ratings, we chose a time threshold of 160 minimum seconds to complete 10 ratings. If a rater was under this threshold (which was based on the average annotation speed of one of the authors), he was automatically discarded. Moreover, we set a maximum of 20 judgments per rater to avoid any tiredness effects. Furthermore, we chose the settings as to prefer quality over speed when recruiting the raters. In addition, we asked 6 expert phoneticians to annotate a subset of the data in order to make sure that their distribution of ratings did not differ significantly from the crowdsourced raters; which it did not. All in this paper discussed analyses are based on the crowdsourced ratings alone.

Raters were instructed to listen to pairs of short audio files (each one $<$10 seconds) and determine in which soundfile a listener sounds more attentive. We defined an attentive listener as someone who a) is paying attention; listens carefully; is observant b) is careful to fulfill the needs or wants of the speaker; is considerate about the speaker.

In a dropdown menu we provided them with a choice to indicate in which file they perceived the listener to be more attentive, but also to indicate when they could not hear any difference in the level of attentiveness.

5. Results

5.1. Thresholds for the Perception of Differences in Attentiveness

In the following subsection we are going to investigate inter-rater agreement depending on prosodic differences in backchannel pairs. For this we divided the ratings into three groups; little, median, and strong agreement. Little agreement consists of all cases in which the winning backchannel only won with a majority of 0-4 votes. Median agreement in turn consist of all cases with a majority margin of 5-9, and strong agreement consists of all cases with a majority margin of 10-23. We investigated both bi- and monosyllabic backchannel token. The results are summarised in Figure 1.

For bisyllabic backchannel token, differences in the duration of backchannels, appear to be a strong cue for raters when deciding which backchannel token conveys more attentiveness. For the second syllable for instance, durational differences of more than 80 ms are more frequent in the "strong agreement" (20%) than in the "little agreement"' group (8%). The opposite trend can be observed for the first syllable. As with duration, rms-intensity appears to be a strong cue. The greater the difference in loudness between the two backchannel token, the more the raters agreed. For instance, for the first syllable, the occurences of differences of 0-0.5 decreases from 37% in the "little agreement"' condition, to 18% in the "strong agreement" group. For the second syllable they decrease from 35% to 19% . We could not find any effect of F0-mean or F0-slope on inter-rater agreement.

For monosyllabic backchannel token, as for the bi-syllabic backchannel token, rms-intensity appears to be a strong cue. Also in this case, the greater the difference in rms-intensity, the more the raters agree. However, different from the bisyllabic backchannels, duration does not appear to be as important, as no clear pattern can be observed. The same is true for f0-slope. In contrast to the bisyllabic backchannels, raters appear to use f0-mean as a cue for their ratings; here the number of instances of a differences greater than 2 increases from 8-24% between little and strong agreement.

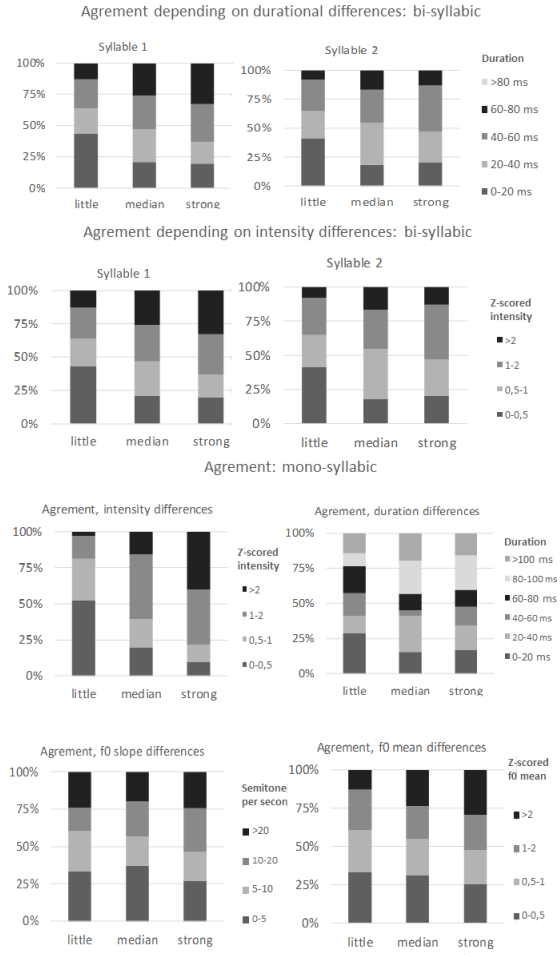

Figure 1: Inter-Rater Agreement Depending on Prosodic Differences.

5.2. More attentive vs. Less attentive backchannels

We made all possible pairwise attentiveness comparisons within the data. We then defined two groups of tokens: “Wi” and “Lo”, in which, for each pair, the more attentive token was assigned to the group “Wi”, and the less attentive was assigned to the group “Lo”. Note that, due to making all comparisons, the groups may contain repeated samples. In addition, we excluded all the backchannel pairs for which the majority of raters indicated that they could not perceive a difference in attentiveness between the two backchannel tokens. The differences across syllable for each prosodic cue between the two groups can be observed for the bisyllabic token in Figure 2 and for the monosyllabic token in Figure 3.

5.2.1. Prosodic Cue Analysis: Bisyllabic Backchannels

Welch Two Sample t-test were conducted to compare duration, f0-mean, f0-slope and rms-Intensity between ”Wi” and ”Lo” stimuli. We found no significant difference in duration for the first syllable. However, there was a significant difference in duration for more–attentive backchannel (M=0.188, SD=0.0524) and less–attentive backchannel (M=0.206, SD=0.0423) conditions; t(476.8)=-4.135, p <0.001. Although there seems to be

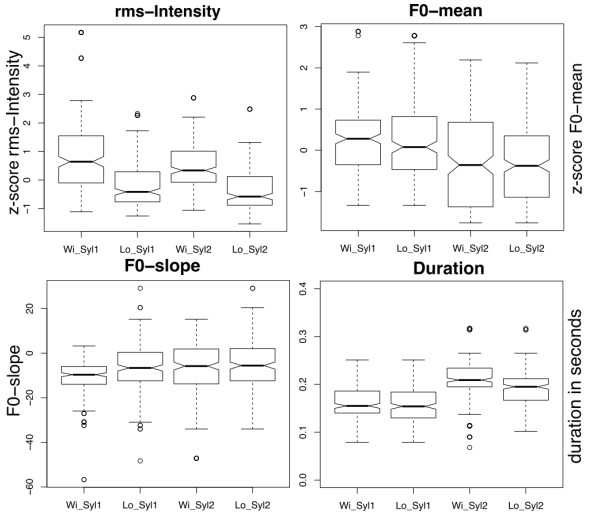

Figure 2: More attentive vs. less attentive bisyllabic backchannel.

a trend that the as more attentive rated syllable has a higher f0-mean, no significant difference in f0–mean for the first or second syllable was found. It is however noticeable that the first syllable is higher in f0-mean than is the second syllable. Moreover, we found a significant difference in F0-slope of the first syllable for more–attentive backchannel (M=-10.516 , SD= 9.3278) and less attentive–backchannel (M=-7.752, SD= 12.907) conditions; t(371.42)=2.485, p = 0.001. Finally, there was a significant difference in rms-Intensity of the first syllable for more–attentive backchannel (M=1.142, SD=1.733) and less attentive–backchannel (M=-0.165, SD=0.849) conditions; t(296.57)= -9.696, p <0.001 as well as the second syllable for more–attentive backchannel (M=-0.309 , SD=0.907) and less attentive–backchannel (M= 0.478, SD= 0.861) conditions; t(406.87)=-9.009 , p <0.001.

5.2.2. Prosodic Cue Analysis: Monosyllabic Backchannels

We carried out the same prosodic analysis also for the monosyllabic backchannel token. We found a significant difference in rms–intensity for more–attentive backchannel (M=0.845, SD= 1.048) and less–attentive backchannel (M=-0.469, SD= 0.793) conditions; t (299.8)= -12.764, p <0.001. The “winning” group has a higher rms-intensity than the “loosing”’ group. We also found a significant difference in f0-slope for more–attentive (M=2.192, SD= 16.552) and less–attentive backchannel (M=-5.643, SD= 15.11092)conditions; t(319.36) = -4.450 p <0.001. No significant difference could be observed for F0-mean and duration.

5.3. RankSVM Classification

In this Section, we evaluated whether it is possible to obtain an automatic assessment on attentiveness.We decided to focus on bisyllabic token in this study, but plan to extend it to include monosyllabic backchannels as well. To avoid defining an explicit attentiveness scale, and stating it as a regression problem, we instead formulated this task as a ranking problem, where two samples are compared according to relative attentiveness. To this end, we employed a Ranking SVM algorithm, in which, for a given pair $\mathbf{x}_i$ and $\mathbf{x}_j$, their difference $(\mathbf{x}_i - \mathbf{x}_j)$ is classified into +1 (if i is more attentive than j) or -1 (if j is more attentive than i), thus turning the problem into a binary classi-

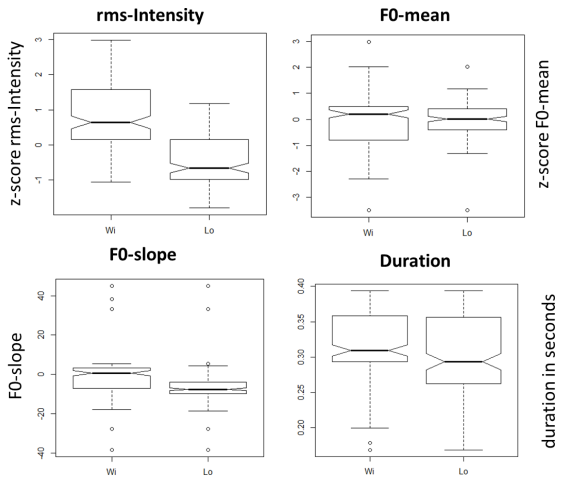

Figure 3: More attentive vs. less attentive monosyllabic backchannel.

fication. To assign such target label (relative attentiveness) to a given pairwise comparison, we used the majority vote from the crowdsourcing experiments. All cases in which the majority of raters indicated that there was no perceivable difference in attentiveness, was excluded from the data set. This reduced the size of the data set to 205 pairs.

In our experiments we used a Support Vector Machine (SVM) classifier based on the linear kernel. We applied a grid-search with 10-fold cross validation to identify the hyper-parameters. For the classification we used all features which were previously determined to have a significant effect on perceived attentiveness except for f0–slope, as this feature reduced the performance of the classifier. The remaining features thus were rms–intensity of first and second syllable as well as the duration of the second syllable. As a result of this experiment, we obtained an accuracy of 83%, which corresponds to a 27% improvement over majority class classification (56% acc.).

6. Discussion

In the current study we have been able to model perceptual differences in mono– as well as bisyllabic backchannel token. It has to be noted, that the bisyllabic backchannel token do not express the prototypical high or rising pitch cues for interest, engagement and surprise. Instead we found that the most salient cues to attentiveness, in these bisyllabic backchannel token, were duration of the second syllable, F0 slope of the first syllable and rms-intensity of both syllables. We did not find any significant difference in F0 mean. For the monosyllabic backchannel token, we found a significant difference in F0-slope as well as intensity. These findings differ from Neiberg et al.[8], who found in their study that surprise and interest are correlated with longer duration and higher average F0. These differences might be due to both, the different characteristica of the respective corpora used, and the difference in approach concerning the distinguishing of bi– and monosyllabic backchannels.

Moreover, previous studies have shown that high or rising pitch is perceived as more engaged than flat pitch (e.g.[15, 16]). However, in the current study we did not find any difference in attentiveness between the bisyllabic tokens with rising or high flat pitch, and those with flat pitch. Furthermore, it has also been found in previous studies that different feedback tokens have different intrinsic functions. In a study on feedback tokens with acted emotions “ah” and “oh”’ was commonly interpreted as surprise, even when the actor tried to convey other functions through prosody [15], and in a study on spontaneous backchannels in casual conversation “aha” and “mhm” have been found to convey surprise [8]. In the current study “aha” and “mhm” were found to differ in how they convey attentiveness. When investigating 71 pairs of bisyllabic feedback tokens of these types that had the same prosodic realization (flat slightly falling pitch) the “aha” tokens where significantly more often regarded as more attentive in 55 comparisons, while the “mhm” token only where rated as more attentive in 12 cases.

As Malisz et al. [9] we found that attentive speakers tend to speak more loudly. While they do not report on any durational aspects and we did not calculate the standard deviation in pitch, our results corroborate their findings in that F0 slope, as a measure of variability, at least for monosyllabic backchannels and in the first syllable of bisyllabic backchannels, is positively related to attentiveness.

7. Conclusion

In the current study we used a corpus of three-party face-to-face groups discussions. We devised perception experiments, and recruited raters through a crowdsourcing platform, in order to investigate how prosodic realisation influence perceived difference in attentiveness in bisyllabic as well as monosyllabic backchannel pairs. To understand, why one backchannel is perceived as more attentive than another, we analysed their patterns of F0, duration and rms–intensity. We found differences in the way attentiveness is expressed in mono–, in contrast to bisyllabic backchannel token. When building a synthesiser for an attentive conversational agent, it might therefore be advantageous, to model mono– and bisyllabic backchannel token separately.

Finally, we have also been able to build a classifier to rank bisyllabic backchannels, that do not exhibit the prototypical high or rising pitched cues for interest, engagement and surprise. The classifier achieves an accuracy of 83%. We believe that this result could further be improved by combining the here proposed classifier with the thresholds for inter–rater agreement discussed in this paper. Such a classifier would only rank a backchannel pair if the difference in prosodic cues extends, a still to be determined, combination of thresholds in prosodic cues. In forthcoming studies, we would also like to use the current findings, as a starting point, in order to devise a parametric speech synthesiser which is able to generate, perceivable more respectively less attentive sounding backchannel tokens.

8. Acknowledgements

The authors would like to acknowledge the support from the Horizon 2020 project BabyRobot (contract no 687831) as well as the Swedish Research Council Project VR(2013-4935)

9. References

[1] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, “Towards robotic assistants in nursing homes: Challenges and results,” *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 271–281, 2003.

[2] M. E. Pollack, “Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment,” *AI magazine*, vol. 26, no. 2, p. 9, 2005.

[3] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[4] K. P. Truong, R. Poppe, I. de Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogs," 2011.

[5] N. G. Ward, "Possible lexical cues for backchannel responses," in *Feedback Behaviors in Dialog*, 2012.

[6] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in *Intelligent Virtual Agents*. Springer, 2008, pp. 176–190.

[7] N. Ward, "Pragmatic Functions of Prosodic Features in Non-Lexical Utterances," *Speech prosody*, pp. 325–328, 2004. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.433

[8] D. Neiberg, G. Salvi, and J. Gustafson, "Semi-supervised methods for exploring the acoustics of simple productive feedback," *Speech Communication*, vol. 55, no. DECEMBER, pp. 451–469, 2013.

[9] Z. Malisz, M. Wodarczak, H. Buschmeier, S. Kopp, and P. Wagner, "Prosodic Characteristics of Feedback Expressions in Distracted and Non-distracted Listeners," *The Listening Talker*, no. May, pp. 36–39, 2012.

[10] T. Kawahara, Z.-Q. Chang, and K. Takanashi, "Analysis on prosodic features of japanese reactive tokens in poster conversations," in *Proc. Intl Conf. Speech Prosody*, 2010.

[11] J. Gustafson and D. Neiberg, "Prosodic cues to engagement in non-lexical response tokens in swedish," in *DiSS-LPSS Joint Workshop 2010, University of Tokyo, Japan, September 25-26, 2010*, 2010.

[12] C. Oertel, K. A. Funes Mora, S. Sheikhi, J.-M. Odobez, and J. Gustafson, "Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews," in *Proceedings of the 2014 workshop on Understanding and Modeling Multiparty, Multimodal Interactions*. ACM, 2014, pp. 27–32.

[13] C. Oertel, K. A. Funes Mora, J. Gustafson, and J.-M. Odobez, "Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 107–114. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820759

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[15] D. Neiberg and J. Gustafson, "Cues to perceived functions of acted and spontaneous feedback expressions," in *The Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Citeseer, 2012, pp. 53–56.

[16] J. Liscombe, J. Venditti, and J. B. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," 2003.