On Data Driven Parametric Backchannel Synthesis for Expressing Attentiveness in Conversational Agents

Catharine Oertel
KTH
Royal Institute of Technology
Stockholm, Sweden
catha@kth.se

Joakim Gustafson
KTH
Royal Institute of Technology
Stockholm, Sweden
jocke@speech.kth.se

Alan W. Black
Carnegie Mellon University
Pittsburgh, United States
awb@cs.cmu.edu

ABSTRACT

In this study, we are using a multi–party recording as a template for building a parametric speech synthesiser which is able to express different levels of attentiveness in backchannel tokens. This allowed us to investigate i) whether it is possible to express the same perceived level of attentiveness in synthesised than in natural backchannels; ii) whether it is possible to increase and decrease the perceived level of attentiveness of backchannels beyond the range observed in the original corpus.

CCS Concepts

•**Human-centered computing** → **Auditory feedback;** Empirical studies in collaborative and social computing;

Keywords

Synthesis; Attentive Agents; Backchannels

1. INTRODUCTION

In recent years more and more research has gone into exploring strategies and possibilities for long term human-agent and human-robot interactions. While most research has emphasised developing strategies for the robot in the role of the speaker, there has also been an array of studies which investigated the role of the listener. Main research questions addressed with regards to the listener role were the development of algorithm for predicting the correct timing of feedback utterances e.g. [6, 14, 7].

In human-human communication it is essential for the speaker to interpret and react to the listeners' nonverbal reactions. Nonverbal reactions can provide information about whether the listener is still interested in what the speaker is saying, or whether he agrees with the speaker, even if this is not expressed through the verbal channel. Similarly also for human-robot or human-agent interaction, the longer a conversation lasts, the more it becomes important for a successful communication to not only provide feedback at the right point in time but also to convey feedback with the correct meaning/attitude. One very important requirement for this to be possible is the availability of a synthesiser which has the capabilities to express meaning/attitude in feedback utterances. Most current speech synthesisers either do not encompass feedback tokens at all, or only support a limited set of stereotypical functions. In order to approach human behaviour, it is important to equip a synthesiser with the same capabilities as a human. This implies that he should be able to express the same variability in feedback token, even if these variations are rather subtle.In the current paper we will address the development of a speech synthesiser which is able to express different degrees of attentiveness in backchannel tokens.

MA3HMI'16 , November 16 2016, Tokyo, Japan
 ISBN 978-1-4503-4562-0/16/11…$15.00
DOI: http://dx.doi.org/10.1145/3011263.3011272

2. BACKGROUND

The two major techniques in speech synthesis at present are unit selection [5] which selects sub-word segments of natural speech; and statistic parametric synthesis [16] which produces a generative model. Each of these techniques has its advantages of quality and flexibility and required amount of data for adequate performance. At first sight it might seem reasonable to use a unit selection technique for synthesis of backchannel utterances, as selecting appropriate natural examples will have high degree of naturalness. This has been done [3], but has the limitation of requiring that the inventory of natural examples must be of significant size to have the desired variance.

Pammi et al. [11] build a synthesiser for listener vocalisations. Their aim was to improve speech synthesis by emotionally colouring listener vocalisations. They chose the best candidate for a given target from among the available vocalisations and then used prosody modification techniques to impose a target intonation contours. They combined markedly distinct intonation contours with vocalisations differing in segmental form, using the prosody modification techniques MLSA vocoding, FD-PSOLA, and HNM. Their findings indicated that the drop in naturalness seems strongest for MLSA and smallest for HNM and FD-PSOLA. They also found that naturalness degrades substantially when imposing intonation contours that are very different from the original contour. They also found unexpected interactions, where certain configurations of segmental form and intonation caused a perceptual impression that was not predictable from the individual meanings of segmental form and intonation separately.

Ward and Escalante-Ruiz [15] implemented a Wizard-of-Oz system in which a tutor was interacting with a student in

a quizz scenario. The authors wanted to investigate in how far alteration in the prosody of the tutor's acknowledgments affected the perception of naturalness and friendliness in the student. More precisely they wanted to investigate whether the students would react more positively if the tutor's acknowledgment matched their previous user state (e.g. that they were sure about the guess, or low in confidence). To achieve this they always used the acknowledgment "Good Job" but realised several alterations of a neutral "Good Job" token from the corpus such as, for example, elongation or creakiness in order to convey praise or expecting the good performance to continue etc. They tested their hypothesis by comparing the students perception after having interacted with the system in comparison to a baseline system. The authors found that the subjects perception of naturalness was significantly higher in their system than in the baseline system.

Stocksmeier et al. [12] used diphone synthesis to produce 12 variants of the German "ja" interjections. They wanted to investigate the influence of prosodic differences on the emotional and pragmatic perception of third party observers. Listeners perceived utterances as bored, hesitant or happy and agreeing depending on the prosodic parameters used for synthesis. They used a spline based pitch curve generator, Ehlich's systematics [4] of the German "hm" as well test recordings of emotional feedback inflection to produce the different variations of "ja". They could show that prosody is an important factor in the perception of emotions such as eg. happy, , hesitant, anxious etc. in German feedback token "ja"'.

Campbell [3] synthesised feedback token using concatenative speech synthesis and for this retrieved situation appropriate token from the large database of the ESP corpus.

2.1 Contributions

The first contribution of the current study is the design and building of a parametric synthesis voice, based on a corpus of on reenacted conversational speech, with a special emphasis on backchannel tokens. The work here has concentrated on using parametric technique, so that we can provide more varied synthetic examples and provide a method of control to produce many targeted appropriate examples.

The second contribution is the preliminary evaluation of perceived attentiveness in synthesised backchannel tokens. While Pammi et al. [11], as well as[12] do investigate the synthesis of emotionally coloured backchannels, they do not investigate the perception of attentiveness in synthesised backchannel token. Moreover, they focus on manipulations of the intonation contour rather than manipulations of intensity or duration. To our knowledge, all other studies concerned with the synthesis of backchannel token focus on the perception of emotions or functions or situation appropriateness. This is the first study which focuses on investigating at what point the significant majority of people perceives a significant difference in the level of attentiveness.

3. DATA

For the following paper we used two corpora. The first corpus was the "KTH–IDIAP Corpus"' [9] and the second one is the "Conversational Synthesis Corpus". The KTH–IDIAP corpus is a corpus of group interactions. A Post-Doc had to identify the best suited candidate for a imaginary prestigious scholarship out of groups of three applicants. In order to achieve as many different conversational dynamics as possible, the recordings were separated into five distinct phases. In the first phase, the three applicants were left by themselves. In the second phase, each applicant was asked to introduce himself in a couple of minutes. In the third phase, each of the applicants had to give an elevator-pitch for the respective research project. In the fourth phase, each of the PhD students had to discuss the potential impact their project could have on society and on the fifth and final phase all three applicants had to come up together with a suggestion for a joined research project. In total, the corpus comprises 5 group interactions of approximately 50 minutes each. The corpus was transcribed and discourse phenomena were annotated.

Figure 1: Setup for KTH–IDIAP Corpus.

Most speech synthesis voices rely on recordings of phonetically and prosodically balanced isolated sentences. Such databases have no examples of discourse functions such as backchannels. Thus we recorded our synthesis training data by placing our voice talent within a dialog context using the transcribed interactions of the "KTH-Idiap corpus" we had our talents (1 male and 1 female) re-perform the texts (with a live dialog partner) so that we could get natural examples of backchannels.The two speakers were seated in two different rooms, but could see each other through a glass wall; thus ensuring no channel bleeding. The whole recording was supervised by a synthesis expert who made sure that the sufficed the quality requirements of speech synthesis. We are aware of others [13] who also record in a dialog situation, but their system was not targeting low level discourse utterances like backchannels. We opted for re-enacting the dialogues as we wanted to same speaker speak all the conversational passages. Using the same speaker has the advantage of not having to account for vocal tract differences when building the synthesis voice and using high-end microphones in a professional recording studio results and better quality recordings. In addition of the conversational passages the speaker recorded, they were also asked to record additional text optimized for synthesis recordings. While reenacting certainly decreases the spontaneity of the data, it provides us with more controlled recordings.

4. BACKCHANNEL SYNTHESIS

In a previous study [10], we found that for bisyllabic backchannel token intensity of the first and second syllable, as well as the duration of the second syllable, is significantly differ-

Figure 2: Setup for Synthesis Recordings.

ent between backchannel tokens which are perceived to be higher in attentiveness from those which are perceived to be lower in attentiveness. We also found that the F0-slope of the first syllable appears to be significantly different in more and less attentively perceived tokens.

For synthesising backchannels, which are different in the perceived degree of attentiveness, we chose to explicitly control for duration and intensity and implicitly control for f0-slope. This means that as we vary the duration, the predicated f0-range may change based on the training data. One reason for this decision is that just knowing that f0-slope differs does not provide sufficient information for prediction. Further information such as the f0-range is needed as well.

We use the Clustergen Parameteric Synthesis System [2] and make use of a random forest based modeling techniques [1], as it performs especially well on limited numbers of examples. For each voice talent we recorded, we only have around 150 instances of backchannels.

The particular set of backchannels we looked at are mostly non-lexical (or at least no clearly articulated as conventional words). Even with words like "okay" there are many examples that do not have clear phonetic articulations of these. Thus we modified our synthesis labeling accordingly. We identified 4 token types ("hmm" (monosyllablic), "mhm" (bisyllabic), okay (bisyllabic), yeah (including variances)). We expanded these with a new "phone" type we called"bc", it is identified as a vowel, and is modeled with three states. Although "bc" is shared between all backchannel types the model is conditioned on the type itself.

Objective measures for synthesis models (for backchannels alone) are in the same space as that for full text to speech for these voices. We get MCD values of 4.54, F0 RMSE values 11.51. These are distortion metrics found by a Euclidean distance between synthesized examples and natural examples. However, the functional evaluation described below are the real measures of adequacy. We quote the object numbers here to show they are not unusual.

We synthesize backchannel examples with a given set of features derived from the original natural examples in KTH–Idiap corpus [9]. These features are z-score normalized for those particular speakers. Likewise, we calculate the same z-scored features for our voice talent rendering of their back channels. These features explicitly control duration and RMS power of the component syllables. The desired features are passed in with the desired token type to the synthesiser and it returns a waveform of that type conditioned on the features. Note that if we just used the token type we would always get the same waveform synthesised.

We do not (at present) have an input feature that explicitly controls the F0 of the backchannel, but as the input features vary, the generated F0 varies as do other aspects of the synthesis (articulation etc). Thus we get variation in all aspects of the synthesised backchannel depending on the z-score features we submit as input.

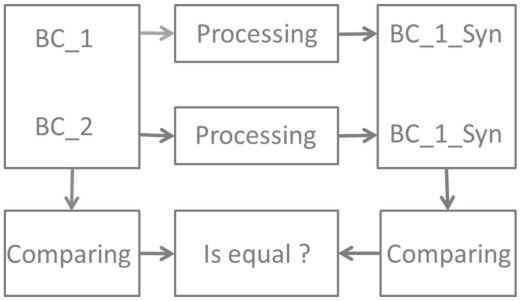

Figure 3: Experimental Flow from sponatneous backchannels to synthesised backchannels.

5. PERCEPTION TEST SETUP

Perception stimuli. Experiments were conducted using the Crowdflower crowdsourcing platform. It consisted of comparisons of the attentiveness level of two feedback realizations. In each case, a carrier sentence from the KTH–IDIAP corpus was used, and synthesised backchannels were inserted at the same place at which a backchannel had occurred in the original recording.This was done to ensure that backchannels were rated in the same interactional environment. We chose carrier sentences which were short in duration, so that we could ensure that the third–party annotators could still remember the previous backchannel when comparing the two tokens.

Ratings. Raters were recruited from the United States, Netherlands, and Germany. They were instructed to listen to the synthesised backchannel pairs and determine in which audio file the listener sounded more attentive. An attentive listener had been described to raters as someone who a) pays 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, raters could indicate in which audio–file they perceived the listener to be more attentive or when they could not see any difference. Also, they could report if the video files did not play correctly in their browser.

To ensure that we received the best quality ratings, we chose a minimum time threshold of 160 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 so as to avoid any tiredness effects. Furthermore we chose the crowdflower settings as to prefer raters with high quality records. Each pair of videos to be compared were annotated by 12 raters.

6. RESULTS

In this Section we report results from the two preliminary experiments investigating the perception of attentiveness in synthesised backchannel token.

6.1 Natural vs. Synthesised Backchannels

The aim of this preliminary experiment was to compare the perceived degree of attentiveness of natural versus synthesised backchannel tokens. The experimental flow of the experiment is illustrated in Figure 3.

Therefore, we first chose 10 natural backchannels (from each 1 male and 1 female speaker of the KTH-Idiap Corpus)

and synthesised them according to their z-scored intensity as well as duration features. The corresponding F0-values were predicted by the Clustergen Synthesiser. We then investigated whether the synthesised backchannels were ranked in terms of attentiveness in the same way as their natural counterparts.

For the female speaker we constructed 25 randomn comparisons. Out of the 25 comparisons there were 9 cases in which a 2/3rd majority was obtained for a preference for one specific feedback token.

In 7 out of these 9 cases, the same backchannel was ranked higher in the synthesised version than it was previously in the original version. For the male speaker we as well constructed 25 random comparisons. Out of the 25 comparisons there were 14 cases in which a 2/3rd majority was obtained for a preference for one specific feedback token. In all cases the feedback token which were ranked higher in the natural token also won the comparison in the synthesised token.

While these were very encouraging results, they did not provide us with any information of how much a backchannel should be louder and longer in order to be perceived as more attentive by a significant majority of people. Therefore we devised the next experiment.

6.2 Perceptible Degrees of Attentiveness

Figure 4 illustrates the second experiment.

We used the rankings of the natural backchannel token (as described in [10]) and used the three highest ranked and the three lowest ranked backchannels. We calculated their average of rms-intensity and duration. We then interpolated between these two points. In order to investigate whether we furthermore could extend the range of attentiveness beyond what we observed in the original corpus, we also extrapolated until and alpha of -0.7 on the one end and 2.0 on the other end. We used an alpha step of 0.2 as to reduce the number of comparisons.

We again followed the perception test setup as described in 5 and made sure that all backchannels (interpolated and extrapolated ones) were combined with each other.

In order to determine how many alpha steps are necessary for a significant number of people to perceive one backchannel token as more attentive than another backchannel token we carried out a chi-square test (normalizing for number of comparisons made). Increasing in alpha value in the "to–be–compared–with–backchannel–token", we found that after the 5th alpha step, the feedback token with the higher alpha receives a significantly higher number of votes from people compared the 1st alpha step backchannel token $X2 = 8.78$, $p < .05$. After the 1st step 43% of people thought that the backchannel with higher alpha sounds more attentive, after the second step 33%, after the third step 49%, after the fourth step 49% and after the fifth 81%.

.

7. DISCUSSION AND CONCLUSION

In the current paper we could show that it is possible to build a synthesis voice which is able to express attentiveness in backchannel tokens. We could show that that we can synthesise the same level of attentiveness in backchannel tokens as is perceived in natural backchannels and that we can also increase and decrease the perceived level of attentiveness beyond the range which we observe in the original corpus. However, it has to be noted that the number of different

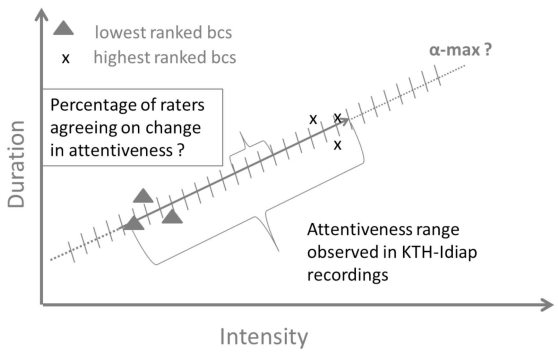

Figure 4: Experimental Flow from sponatneous backchannels to synthesised backchannels.

backchannels we were able to synthesise was constrained by two things. On the one hand, we could not decrease the intensity too much as otherwise it was not possible to hear the backchannel token any more. On the other hand, after an alpha value greater than 2.0, the predicted pitch slope changed and thus also the perceived backchannel function. Within this given scope from alpha -0.7 to 2.0, this meant there were at least 2 distinct levels of attentiveness as well as one possible further level which will be perceived by still a fair number of people.

While these preliminary results are certainly encouraging they are of course limited in that we are only synthesising and investigating the backchannel token "mhm". In order to test the generalisability of this approach it would be also important to test it on other feedback token, such as for example "okay" or "yeah". In future work we will furthermore extend the capabilities of the synthesis to also include other paralinguistic phenomena such as for example certainty, agreement and disagreement.

In a previous study [8], we could show that third party observers can distinguish between three distinct listener categories, "attentive listener", "side–participant", "bystander", on the same corpus. We could furthermore show that gaze patterns, as well as the frequency of head nods, and backchannels are significantly different between the different listener categories. Currently, we are working on bringing these studies together and to implement an attentive listening agent who is able to express the appropriate degree of attentiveness appropriate for the different listener categories.

We are planing to release the synthesis voice by the end of this year.

8. ACKNOWLEDGMENTS

Catharine Oertel and Joakim Gustafson 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] A. Black and P. Muthukumar. Random forests for statistical speech synthesis. In *Interspeech 2015*, Dresden, Germany, 2015.

[2] A. W. Black. Clustergen: a statistical parametric synthesizer using trajectory modeling. In *Interspeech 2006*, 2006.

[3] N. Campbell. Towards conversational speech synthesis; lessons learned from the expressive speech processing project. In *SSW 2207*, pages 22–27, Bonn, Germany, 2007.

[4] K. Ehlich. Interjektionen. *Max Niemeyer Verlag*, 1986.

[5] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP-96*, volume 1, pages 373–376, Atlanta, Georgia, 1996.

[6] B. Inden, Z. Malisz, P. Wagner, and I. Wachsmuth. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 181–188, 2013.

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

[8] 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 *International Conference on Multimodal Interaction*. ACM, 2015.

[9] 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*, UM3I '14, pages 27–32, 2014.

[10] C. Oertel, J. Gustafson, and A. W. Black. Towards building an attentive artificial listener: On the perception of attentiveness in feedback utterances. In *Proc. of Interspeech*, pages 2915–2919, 2016.

[11] S. C. Pammi, M. Schröder, M. Charfuelan, O. Türk, and I. Steiner. Synthesis of listener vocalisations with imposed intonation contours. In *Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*. ISCA, ISCA, 2010.

[12] T. Stocksmeier, S. Kopp, and D. Gibbon. Synthesis of prosodic attitudinal variants in german backchannel ja. In *Interspeech 2007*, pages 1290–1293, Antwerp, Belgium, 2007.

[13] A. Syrdal, A. Conkie, Y. Kim, and M. Beutnagel. Speech acts and dialog tts. In *SSW 7*, Keihanna, Japan, 2010.

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

[15] N. G. Ward and R. Escalante-Ruiz. Using responsive prosodic variation to acknowledge the user's current state. In *Interspeech 2009*, Brighton, UK, 2009.

[16] H. Zen, K. Tokuda, and A. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1059–1064, 2009.