The effect of prosodic features on the interpretation of synthesised backchannels

Åsa Wallers, Jens Edlund, and Gabriel Skantze

Department for Speech Music and Hearing, KTH
Lindstedtsv. 24, 100 44 Stockholm, Sweden
{wallers, edlund, gabriel}@speech.kth.se

Abstract. A study of the interpretation of prosodic features in backchannels (Swedish /a/ and /m/) produced by speech synthesis is presented. The study is part of work-in-progress towards endowing conversational spoken dialogue systems with the ability to produce and use backchannels and other feedback.

1 Introduction

Spoken dialogue among humans is an intricate and fine-tuned process which puts high demands on the participants' ability to perceive and produce inputs and outputs according to the flow of the dialogue, as well as to the context. In a conversation, the participants take turns talking, and the speaker transition is for the most part a very smooth interaction with little speech overlap [1].

Interaction control in spoken dialogue systems is an active area of research. We are becoming increasingly good at dealing with online analysis of human speech and great efforts have been spent to make systems give properly timed feedback. Many researchers working with the development of spoken dialogue systems have shown interest in prosodic features when trying to make the system handle the turns properly in the conversation (e.g. [2, 3]).

As our research systems become more human-like and better at timing their responses, other shortcomings become more apparent. In human-human dialogue, feedback and back-channels make up a significant part of the interaction, and a spoken dialogue system that is to be deemed responsive and human-like needs similar capabilities. We have made preliminary user studies indicating that backchannels have a great effect on how a conversation proceeds, and similar observations are described in more detail by Riccardi and Gorin [4].

A problem that has to be overcome in order to achieve system backchannels is that the interpretation of feedback backchannels, such as *ah* and *m*, may depend on their prosody. In the type of unrestricted conversations we are aiming at in our research systems (e.g. Waxholm, August, AdApt [5], and currently Higgins [6]) the demands on flexible output generation makes canned speech difficult to use. Instead, we aim to include prosodic variation in synthesised feedback and backchannels.

2 Method

Previously we looked at the effect of prosody on one-word elliptical feedback [7]. Here, we attempt to the same with the Swedish one-syllable back-channels *m* and *a*.

These are commonly used in Swedish: in Swedish MapTask dialogues [8], we found that *a* and *m* made up 34% and 15% of the backchannels, respectively.

The stimuli consisted of monosyllabic renditions of /a/ and /m/ synthesised using an experimental version of LUKAS diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool. Although disyllabic /mm/ and /aa/ also occur frequently in Swedish, we used the monosyllabic versions only, partly to constrain the dimensionality of the experiment and partly in an attempt to make the experiment consistent with [7].

For each of the two test words the parameters peak POSITION, peak HEIGHT, and DURATION were manipulated. Three peak positions were obtained by time-shifting the focal accent peaks in intervals of 100 ms comprising EARLY, MID and LATE peaks. The LOW and HIGH peak height was set to 130 and 160 Hz. The durations SHORT and LONG were set to 450 and 650 ms. Combination of the two backchannels and the three properties gave a total of 24 different stimuli, schematically represented in Fig. 1.

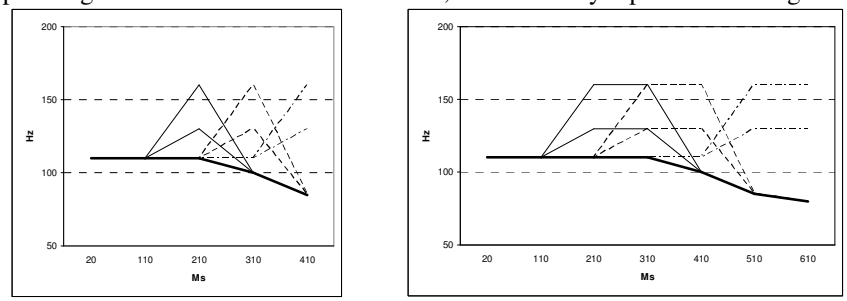


Fig. 1. The prosodic properties of the short and long stimuli, respectively.

A preliminary listening test made it clear that the results found in [7] (i.e. a mapping between peak position and level of grounding) would not be reproduced on these stimuli. Subjects' comments showed a wide range in interpretations of the stimuli, so we resorted to a two-step exploratory approach.

Table 1. List of backchannel interpretations.

Interpretation (in English translation)	Abbreviation
Good, you are in the right place.	RIGHT PLACE
Oh, NOW I understand where you are.	OH!
Really? That was unexpected.	UNEXPECTED
Oh, you are in the wrong place.	WRONG PLACE
Okay, but I need more information.	CONTINUE

In a first experiment, five listeners were subjected to dialogue fragments consisting of a human speaker uttering "On my left I have a X house…", where X was one of the colours red, yellow, and blue, followed by one of the system backchannels. This was repeated for each stimuli and colour, and after each fragment, the listener was asked to write down a free interpretation of the system's response. These interpretations were then manually summarised and condensed into the five paraphrases found in Table 1.

The second step was a perceptual test where the eight participants were asked to listen to the backchannels in the context, and then chose the one of the five paraphrases they felt best represented the meaning of the backchannel. Each stimulus was played three times and the order was randomised.

3 Results

In general, the results show that the variation in prosodic features does effect the interpretation of the backchannels significantly, and in general, the choice of stimulus (/m/ or /a/) had a greater effect than anticipated, with /a/ tilting the interpretations heavily towards OH! and /m/ showing a preference for CONTINUE (Fig. 2). Unfortunately it is difficult to draw any general conclusions about the individual effect of each parameter – POSITION, HEIGHT, and DURATION – from the material. The exploratory experiment design makes it difficult to test for significance, and as this is work-in-progress, we will limit the presentation to raw numbers.

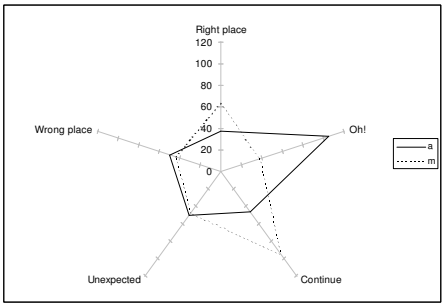
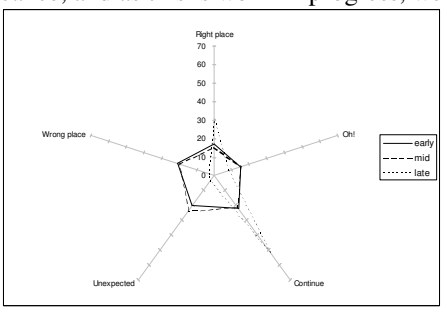
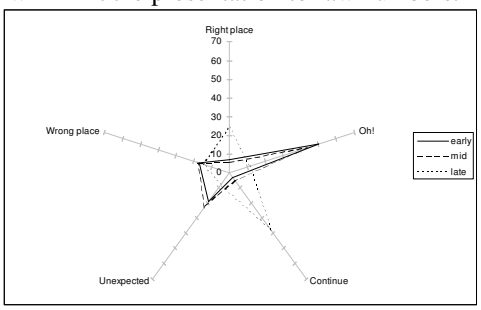


Fig. 2. Distribution of votes for /a/ and /m/





Figs. 3 and 4. Distribution of votes for different setting of PEAK for /m/ and /a/, respectively.

The feature that shows the greatest difference in numbers is peak POSITION. EARLY and MID position give similar results both for /a/ and /m/, but LATE leads to a different interpretation. In the case of /m/, the interpretation of EARLY/MID POSITION stimuli is very vague, with close to equal distribution amongst the five interpretations (Fig. 3). For /a/, the same features leads to a bias towards the OH! interpretation (Fig. 4). For /m/ and /a/ alike, a LATE peak position shifts the bias heavily towards the more neutral CONTINUE interpretation (Figs. 3 and 4). The HEIGHT and DURATION parameters show less clear influence on the distribution of interpretations.

The stimuli reaching the highest consensus amongst the subjects are /a/ LONG EARLY HIGH peak and /a/ LONG MID HIGH peak, where the OH! Interpretation obtained 75% and 71% of the votes, respectively. For /m/, LONG and SHORT LATE LOW peak yielded the highest consensus, with 63% and 58% of the votes, respectively, for CONTINUE. Finally, for both /m/ and /a/ RIGHTPLACE obtained 46% of the votes in the SHORT LATE HIGH peak setting. The full results are available in detail in [9].

4 Discussion and future work

The preliminary work described here has taught us valuable lessons:

- it is indeed possible to produce synthesized backchannels with variable prosody that is perceived and interpreted in a consistent manner by human subjects
- our previous findings on the interpretation of prosodic patterns applied to synthesised monosyllabic one word clarification ellipses ([7]) are not directly applicable on synthesised monosyllabic backchannels
- peak POSITION may effect the interpretation of monosyllabic backchannels more than DURATION and HEIGHT

As the study is context dependent, it needs expansions in several ways to test its relevance to a wider domain. Presently we intend to examine if the backchannels are sufficiently salient to signal system reactions in a spoken dialogue system, as well as to study how system bachchannels are perceived in general is such a setting.

Finally, we made several observations that beckon further investigation. For example, the interpretation categories derived from the open interpretations may be grouped into RIGHTPLACE, WRONGPLACE and CONTINUE on the one hand, and OH! and UNEXPECTED on the other. The first group combine interpretations suggesting that nothing unexpected has occurred from the point of view of the producer of the backchannel, whereas the second group combines interpretations that include an element of surprise.

Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations and was supported by the EU project CHIL (IP506909).

References

1. Levinson, S. (1983). Pragmatics. Cambridge: Cambridge University Press.
2. Edlund, J & Heldner, M (2005). Exploring Prosody in Interaction Control. Phonetica, 62(2-4).
3. Ferrer, L., Shriberg, E., Stolcke, A (2002).: Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. Proc. Int. Conf. spoken Lang. Processing, Denver, pp. 2061–2064.
4. Riccardi, G. & Gorin, A.L. (2000): Spoken Language Adaptation over Time and State in a Natural Spoken Dialog System. IEEE Trans. on Speech and Audio.
5. Gustafson, J. (2002): Developing Multimodal Spoken Dialogue Systems. Empirical Studies of Spoken Human-Computer Interaction. TRITA-TMH 2002:8, ISSN 1104-5787.
6. Edlund J, Skantze G & Carlson R (2004): Higgins - a spoken dialogue system for investigating error handling techniques. In Proceedings of the International Conference on Spoken Language Processing (ICSLP) 2004 (pp. 229-231). Jeju, Korea
7. Edlund, J, House, D, & Skantze, G (2005): The Effects of Prosodic Features on the Interpretation of Clarification Ellipses. In Proceedings of Interspeech 2005, Lisbon, Portugal
8. Helgason, P. (2002). Preaspiration in the Nordic languages: synchronic and diachronic aspects. Doctoral dissertation, Stockholm University, Stockholm.
9. Wallers, Å. (2005): Minor Sounds with Major Impact. Master thesis, KTH, Stockholm.