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](image)

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

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