A corpus based analysis of back-channel vocalizations

Sathish Pammi and Marc Schröder
DFKI GmbH, Saarbrücken, Germany
{Sathish.Pammi@dfki.de, Marc.Schroeder@dfki.de}

Introduction

Back-channel vocalizations play an important role in communicating listener intentions while the other person has the turn or other is talking. The communicative intentions behind back-channels not only transmit messages like 'I am listening' and 'I am with you', but also transmit listener affective states like excited, bored, confused, surprised, and so on. Synthesis of back-channel vocalizations is one of the focused research areas to improve emotionally colored conversational synthesis, and includes different research questions like where to synthesize, what to synthesize and what kind of acoustic properties have to be obeyed to communicate different affective states in different situations. Already a few attempts were made in this area of research; for example, the importance of affect bursts as a feedback in a conversation was investigated (Schröder et al, 2006) through listening tests, Nigel Ward and Wataru Tsukahara (2000) had developed some rules to generate back-channel responses in a conversation and investigated how to use low pitch regions as cues for back-channel responses. However, the analysis and identification of distinguishable types among back-channel vocalizations, their acoustic properties and affective states behind them have to be studied as they are crucial to improve interactive speech synthesis.

This extended abstract explains about a method for collecting back-channel vocalizations and our ongoing work on annotation and a simple data and acoustic analysis of these vocalizations.

Method for database collection

Traditionally, speech or expressive speech synthetic databases were recorded in studio environment with a single speaker or an actor. But when we look at back-channel vocalizations, they appear natural only in conversation. Considering the above issues, we opted to record dialog speech in a studio environment. When a professional German actor was engaged in a conversation, recordings were made from different channels in sessions of about 20 minutes each. The instructions were given to the actor to keep the conversation live as long as he can act like a specific character among Spike, Obadiah, Poppy and Prudence. Each character represents different emotional states: Spike is always aggressive, Obadiah is always gloomy, Poppy is always happy and Prudence is always neutral. Our student assistants, acting as dialogue partners, tried to keep the actor in listening mode for a maximum amount of time while they were interacting with the actor on a specific topic. The speakers were sitting in separate rooms but saw each other through a glass wall. Each speaker's voice was recorded on a separate channel.

Methods for annotation

From the first sight, when we look at the interactive speech corpus, we observed that many of the non-verbal vocalizations made by the actor belong to three broad categories: back-channel, affective and laughter vocalizations Different types of non-verbal vocalizations like affective back-channels, laughter as back-channels and affective laughters like amused laughter were observed. So, an ABL annotation schema was proposed to annotate this kind of data, where A stands for Affective, B stands for Back-channel and L for Laughter. The corpus was annotated by two student assistants according to ABL-schema using Praat software. To annotate the turn or 'floor' of the actor automatically for reducing some efforts in manual annotation, a simple algorithm was developed, allocating turn based on energy. We annotated only non-verbals produced by our target speaker, not of the interaction partner.

Results

As a result of efforts made in database collection and annotation, we obtained six hours of German dialog speech with ABL annotation. We identified 1175 non-verbal vocalizations, among them 918 non-verbals (78%) were noted as back-channels. Among all back-channel vocalizations, 298 (32.4%), 68 (7.4%) and 38 (4.1%) back-channels are noted as Affective, Laughter and Affective-Laughter respectively.

We identified that the actor had spent most of the time in listening mode, and the actor's recording time spent as speaker and listener was 32% and 68% respectively. Around 33% of back-channels were labeled as affective, that means that one third of the back-channel vocalizations were transmitting affective states through them. Interestingly, 3% of non-verbal vocalizations are annotated as laughter, but not as affective or back-channel. It could be interesting to find the meaning or functions behind this kind of vocalizations through informal description, which will be available at later stages.

For a first overview, we analyzed some acoustic properties of back-channel vocalizations made by different characters. The average duration of back-channels provided by Spike, Prudence, Poppy and Obadiah were noted as 0.79, 1.26, 0.98, 0.66 and 1.74 seconds, respectively. On average, Obadiah and Prudence were producing back-channel responses 4.8 times per minute, whereas Poppy is producing only 2.7 back-channels per minute (shown in Figure 1). Another observation is that many back-channel responses given by Obadiah are unvoiced responses and nasal sounds like mhm, hmm or hhh (Figure 1). Other acoustic features also show systematic differences. For example, the mean pitch (F0) of voiced segments in Poppy is high compared to other characters (not shown in Figure).

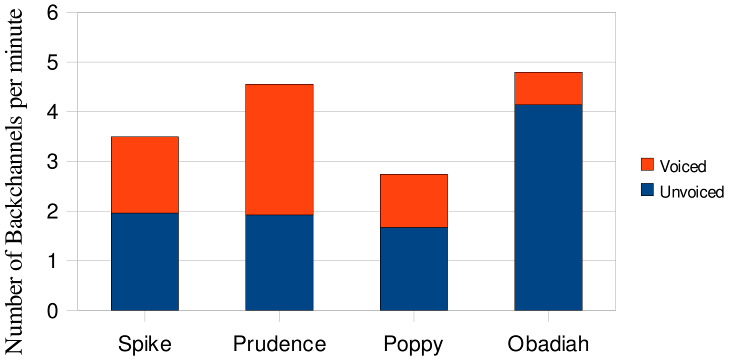

Figure 1: Average number of back-channels produced by different characters per minute and percentage of voiced-unvoiced vocalizations

When we conclude altogether, a useful observation in terms of interactive speech synthesis is that the gloomy character (Obadiah) produces an average of 4.8 back-channels per minute, most of them nasal sounds with long durations around 1.74 seconds, whereas our happy character (Poppy) utters only 2.7 back-channel responses per minute, which are relatively short utterances around 0.66 seconds.

To understand better the structure of both behavior and function of non-verbal vocalizations, we are currently annotating all non-verbals using informal descriptions. Subsequent clustering of these descriptions will help understand the types of form and meaning of non-verbal vocalizations, at least for the speaker we studied.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE). We would like to thank Professor Roddy Cowie and Dr. Dirk Heylen for useful discussions.

References

Schröder, M., Heylen, D. and Poggi, I. (2006). Perception of non-verbal emotional listener feedback. *Proc. Speech Prosody 2006,* Dresden, Germany.

Nigel Ward and Wataru Tsukahara (2000). Prosodic Features which Cue Back-channel Responses in English and Japanese. *Journal of Pragmatics*, 23, pp 1177--1207