**BAS Infrastrukturen zur Technischen Sprachverarbeitung (BITS)**

**Teilprojekt Nr 8 (Doku 8/1e)**

# *Protocol for the Casting*

**Content:** Below the casting of the speakers for the BITS project is described. The selection of the speakers and the recording conditions are explained. In the results some items are included which should be considered in future recordings.

**Author:** Tania Ellbogen

**Date:** 17.10.2005<br>
**Version:** 1.1

## Protocol for the casting

It was planned to record a diphone corpus for speech synthesis for German which includes every German phoneme plus seven English and three French phonemes. The phonemes should be recorded in every diphone combination that is possible in German. Before these recordings were made we recorded a mini corpus in the studio of the IPSK. This mini corpus should comprise the required diphones for the synthesis of three German sentences. The sentences were selected under the aspect to include as many different German phonemes as possible. The selected sentences were: „Heute ist schönes Frühlingswetter.“, „Wer muss noch Schularbeiten machen?“and „ Der herrische Pate versteht sich als Pol der ganzen Familie.“ The diphones were embedded in logatomes with the diphone as part of the second or third syllable.

Speakers:

We invited 45 speakers who had to read a list of 90 logatomes one by one from a screen. Additionally 19 of the 45 speakers had to read a list of 90 reasonable German words. The age of the speakers ranged from 12 to 72 years (two of them were children). The speakers lived in south German regions predominantly. Most speakers (33 out of 45) were professionals or actors with trained voices.

Duration of the recordings:

The speakers were invited one by one to the recording sessions. The recordings were done over several weeks in winter 2002/03 and took about half an hour each.

Recording procedure:

The casting took place in the studio of the IPSK. The speakers were seated in a room with low reverberation. The general speaking direction of the speaker is at a light angle to the only window surface to prevent direct echoes. The signals were recorded on two channels: close talk microphone (Beyerdynamic NEM 192) positioned 7 cm to the left of the midsagittal plane at the height of the upper lip, and laryngograph signal (LaryngoGraph PCLX).During the session the speech prompts were displayed through the window on a screen outside the recording room.

Two supervisors outside the recording room monitored the recording: a controller provided the prompts and one person was responsible for that the pronunciation uttered by the speaker was as intended. Whenever there was a change of quality of the voice or the speaker needed to there was a break.

Results:

- Despite two supervisors a consistent quality of every logatome and word could not be achieved.

- In the synthesis it turned out that the quality got worse the slower the prompts were spoken.
- Noticeable gaps in the words of the synthesised sentences arose from the fact that the diphone was separated by two syllables.
- The synthesised sentences out of reasonable German words are spoken in considerably deeper pitch of voice than the synthesised sentences out of logatomes. The exact reasons for this could not be determined until now but there are two hypotheses:

  1. If the speakers read logatomes they are under greater stress than if they read known words. By  production of adrenaline a higher pitch of voice is evoked.

  2. „If a speaker is to read logatomes it is impossible for him to fall back on learned prosodic structures. Thus he is forced to structure the logatomes by himself. This is usually achieved with prosodic elements like variation of dynamic, tempo and voice pitch. Furthermore there is the need of the speakers to give sense to the logatome. So the intention of the speaker is important on how he pronounces a logatome.“ (Dr. Ulla Beushausen)