The BITS Speech Synthesis Corpus for German

Tania Ellbogen, Florian Schiel, Alexander Steffen

Bavarian Archive for Speech Signals (BAS)
c/o Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität, Schellingstr. 3, 80799 München, Germany
{ellbogen, schiel, al-x}@bas.uni-muenchen.de

Abstract

In this paper we announce the new BITS[1] Synthesis Corpus for German. The BITS project is funded by the German Ministry of Education and Science to provide a publicly available synthesis corpus for German. The corpus comprises the voices of four German speakers (two male and two female) and consists of two parts: a set of logatome recordings for controlled diphone synthesis and a set of sentence recordings for unit selection. The paper gives an overview about the basic specifications, the profiles of the speakers, the casting procedure and quality control. Annotation and its organisation are described in detail. The final BITS speech synthesis corpus will be available via BAS and ELDA probably end of 2005.

1. Introduction

Speech synthesis using concatenative techniques is maturing to a point where standard procedures are being implemented in a variety of products. However, because of the considerable costs most small and medium-sized companies as well as university labs cannot afford to produce the required speech resources on their own. Although there are some public domain German diphone voices available for research purposes (e.g. MBROLA) [1] there is definitely a lack of publicly available synthesis resources. Therefore the Bavarian Archive for Speech Signals (BAS) applied for public funding to close this obvious gap within the BITS project ([2]) funded by the German Ministry of Education and Science. According to the project plan the release of the final resource is scheduled for the end of 2005.

In 2002 BAS invited a group of international experts working in the field of speech synthesis to Munich for discussions about the properties of the planned speech resource. Since then the group in BITS working on this project has achieved considerable progress. In this article for the first time we will report about the recording, labelling and the properties of the speech corpus produced. We would also like to invite interested scientists and application engineers to provide additional input, comments, proposals on how to annotate and enrich the basic recordings for practical applications that we might not have included in our specs up to now.

In the following we will give an overview about the specification – especially the selected contents of the two recording sets -, describe the speaker casting and the procedure that led to the final four voices, the recording procedure and annotation, and finally give some information about the quality control and availability.

2. Specification of the synthesis corpus

The synthesis corpus consists of two parts: a set of logatome recordings for controlled diphone synthesis and a set of sentence recordings for unit selection techniques. All sentences and logatomes have been transcribed in SAM-PA. We found SAM-PA a suitable alphabet for speech synthesis because different variants of pronunciation for one phoneme are comprised in one code similar to the orthographic form of a word, and on the other hand it is no problem to encode French and English phonemes. Note that we use here the extended German SAM-PA alphabet as being used in all BAS projects (e.g. [9]).

2.1. Logatome Set

Because of the numerous English and French words (movie titles, names of restaurants etc.) used in daily spoken German we thought it essential to include some English and French phonemes.

Therefore we extended the basic German SAM-PA set of 45 German phonemes (/I/, /E/, /a/, /O/, /U/, /Y/, /9/, /i:/, /e:/, /E:/, /a:/, /o:/, /u:/, /y:/, /2:/, /aI/, /aU/, /OY/, /@/, /6/, /?/, /p/, /b/, /t/, /d/, /k/, /g/, /pf/, /ts/, /tS/, /f/, /v/, /s/, /z/, /S/, /Z/, /C/, /x/, /j/, /h/, /m/, /n/, /N/, /l/, /R/) by seven English (/EI/, /@U/, /T/, /D/, /r/, /L/, /w/) and three French phonemes (/E~/, /a~/, /o~/). Without the latter a realistic speech synthesis for German would not be feasible.

We used these phonemes to generate 2783 diphones that are embedded into an articulatorily neutral context: /a/ or /@/ for consonants, /t/ or /d/ for vowels, for example:

"pat<u>ehpf</u>adau" /e:pf/
"ad<u>eusch</u>adei" /OYS/

Within this neutral context we expected the least coarticulatory effects. The diphones are embedded in logatomes so that the diphone is part of the second or third syllable. The logatomes all end on -au, -eu, -ai and should be read in a monotonous manner with stress mainly on the last syllable. It is essential that particularly the diphone is not stressed because it is easier in speech synthesis to stress a syllable where it is necessary than to remove the stress of syllables. The best way to achieve unstressed diphones would have been to embed the logatomes in a carrier sentence (e.g. 'Ich habe patu<u>ck</u>adau gesagt.' *engl.: 'I said patuckadau this time.'*) but for economical reasons we decided against that. Probably the recording time would have been twice or three times as long as it is now. Furthermore it would be too boring for the speakers to repeat the same carrier sentence more than 2500 times.

During the casting we found that the concatenative synthesis gets considerably worse if the phones of the

1 BITS = **B**AS **I**nfrastructures for **T**echnical **S**peech Processing

diphone are separated by a syllable boundary. In this case you often can hear a "break" in the synthesised word which sounds more unnatural than synthesised speech normally does. On the other hand it is impossible in German that all diphones are within one syllable. To circumvent this problem we advised the speakers to read these diphones rather fluently.

For the content of the logatome list we used a set of well known phonological rules to exclude impossible combinations (e.g. no /N/ at the beginning of a word). There were still many diphone combinations left that seemed very unlikely. But since no reliable rule could be found that guaranteed that all possible diphone combinations in German were created and all others were excluded, we went the other way, i.e. all diphone combinations were kept in the set that can be produced by a trained speaker even if there is a high probability they are actually never needed in daily speech. The same applies to combinations with English or French phones. On the other hand we preferred to have too many diphones than too few.

2.2. Unit Selection Set

The set consists of 1683 sentences which were to be read fluently with normal intonation. The sentences are a subset selected from the TAZ corpus (07/01/1988 – 06/30/1994) by a greedy algorithm plus semantically unpredictable sentences that make no sense but are grammatically correct in German (kindly provided by the synthesis group at IMS, University of Stuttgart, Germany). Furthermore the set contains trade names and proverbs. The sentences were selected to cover every possible German diphone combination in as many contexts as possible ([3]).

3. Casting of Speakers

We invited 45 speakers – professionals and nonprofessionals - for a casting. They were asked to read 90 logatomes that contained a subset of our diphone set so that three target sentences covering nearly all German phonemes could be synthesised. The sentences were: „Heute ist schönes Frühlingswetter.", „Wer muss noch Schularbeiten machen?" and „Der herrische Pate versteht sich als Pol der ganzen Familie." These sentences were rated by 18 people (mostly phoneticians) regarding naturalness and pleasantness. Based on this ranking 10 speakers were selected as nominees. Samples of their recordings were sent to a group of international experts in speech synthesis. In an overall evaluation of all inputs as well as from the judgements of the BITS group the best four speakers (two male and two female) were chosen for the final recordings.

Based on a recording pretest we estimated 25 sessions of one hour for each speaker to cover both recording sets. The sessions were divided into two parts. In the first 30 minutes logatomes, in the last 30 minutes unit selection sentences were recorded.

3.1. Speaker Profile

Originally the desired profile of the speakers called for a male and a female between the age of 20 to 30 and a male and a female between the age of 40 to 50. Ideally, all should be professional speakers with German as mother tongue (L1) and with foreign language competence in English and French (L2). Unfortunately we could not find four speakers in the top range of the ranking that fitted all these characteristics. Since the most important attributes are naturalness and pleasantness in synthesised speech we selected the following four speakers (see table 1).

	spk1	***spk2***	***spk3***	***sp4***
sex	f	f	m	m
age	47	45	40	38
smoker	+	+	+	-
L2	E, F	E	E, F	E, I
years of training	1	-	3	3
profession	Radio announcer	Adviser Painter	Radio announcer	Actor

Table 1: Profiles of the four selected BITS speakers

All four speakers and their parents have German mother tongue, all of the speakers live in Bavaria, Germany.

Although it is possible that the voices of untrained laymen or semi-professional speakers are as good as voices from professional speakers we strongly recommend that only professionals are recorded. It turned out that working with semi-professional speakers is very exhausting for the speaker as well as for the recording staff. Semi-professional speakers can get bored or even angry when they are asked to read meaningless phrases for extended periods, and even worse, they can have great difficulties in pronouncing the words correctly. Besides, you will need a lot more time for recording and therefore more money. The only semi-professional speaker in the BITS speech recordings needed about twice as much time as the other speakers. Therefore we decided after five sessions to record exclusively the logatome set for this speaker to keep our costs in line with the budget.

Another point is that the recordings should not take place early in the morning. Professional speakers are unanimous that the voice needs some wake time to sound at its best.

For future castings we recommend making sure that the speakers are able to pronounce every necessary phoneme correctly. It turned out that some of the BITS speakers have great difficulties producing voiced vs. unvoiced (e.g. /s/ vs. /z/) or non-German phonemes.

4. Recording Procedure

The speaker is seated in an insulated room with low reverberation. The room is acoustically de-coupled from the rest of the building to dampen the background noise. The general speaking direction of the speaker is at a light angle to the only window surface to prevent direct echoes. The speaker is told to adopt a comfortable position and not to move during the recordings if possible. The speaker

is given a beverage and is reminded to ask for a break whenever she needs to.

The positions of the chair and room microphone are marked on the floor. Furthermore we noted the angle of the rack of the microphone to the wall. In this way it should be guaranteed that speaker and equipment have the same position throughout all sessions. If the speaker declares herself indisposed before the recording (e.g. because she has a cold), the recording session is cancelled.

Before each recording session the placements of microphones and the laryngograph electrodes are checked (see figure 1) and the sound signals and the lanrygograph signals monitored on a oscillograph. A check list is run over all settings of the sound mixer and the PC software.

During the session the speech prompts are displayed through a window on a screen outside of the recording room. Three supervisors outside the recording room (see fig. 1) monitor the recording: a controller provides the prompts and listens for technical noises, one person pays attention to the correct speaker intonation and one person is responsible that the pronunciation uttered by the speaker is exactly according to the desired specifications of the prompt. The recording of a prompt is repeated until all three supervisors give their consent. In case of the logatome set the pronunciation has to be exactly canonical and no deviation is tolerated. Furthermore the diphones (and therefore the logatomes) should sound "natural", i.e. they have to be spoken fluently.

Fig. 1 : Monitoring of a recording. In the background the window to the recording room is visible.

Unit selection sentences should be read canonically but fluently as in high quality readings or radio announcements. Phenomena like schwa-elision are natural in German and therefore tolerated. In words or names that can have different possible variants of pronunciation we accept both (or more) pronunciation forms.

The complete recording session is handled by *SpeechRecorder*, a software package developed within the BITS project (see [4] in this conference). *SpeechRecorder* uses a dual head display mode: The monitor visible to the speaker displays only one logatome or sentence at a time while the controller can see the prompt specifications, a level indicator and the last recorded sound wave. The controller can move forwards and backwards through the list of prompts. Aside from the direct sampling into the connected PC each session is recorded on a DAT cassette and after each session the recorded data are saved directly to an external server running RAID 5. For security reasons all data are transferred every night to a server at a different location on the campus.

5. Technical Specifications

The speech signals are recorded with a sampling rate of 48 kHz, 16 bit via a Yamaha O2R digital sound mixer directly to hard disc using the multi-channel recording software *SpeechRecorder* ([4]):

- Channel 1 : close talk microphone (Beyerdynamic NEM 192) positioned 7cm to the right of the mid-sagital plane at the height of the upper lip.
- Channel 2 : large membrane condenser microphone (Neumann Type TLM 103) 60cm from the mouth.
- Channel 3 : laryngograph signal (LaryngoGraph PCLX)

Channels are separated into standard WAV format files; no further processing is performed to avoid any undesired degradations of the signals.

6. Annotation

For the phonetic annotation all logatomes and sentences will be segmented in a first pass with MAUS ([5]) into German SAM-PA. MAUS automatically produces phoneme segmentations for words or whole sentences either by forcing the labelling to follow the canonical pronunciation (forced alignment) or by using a stochastic model of possible pronunciations to a given utterance and then produce the most likely pronunciations to the recording (in the following referred to as "MAUS version").

In a pre-test we evaluated which of both techniques is more suitable for the logatome and sentence segmentation respectively: four trained phoneticians segmented the same three sentences of each speaker that were pre-segmented both in the canonical form and in the MAUS version. After that they decided together which of the versions they preferred. It turned out that for the unit selection sentences the MAUS version was more efficient to work with (probably because of the typical German pronunciation phenomena already covered by MAUS).On the other hand the logatomes are pre-segmented according to their canonical form. This guarantees that the logatome contains the diphone in correct SAM-PA transcription which may not be the case in the MAUS version. Since we are not interested in the reminder of the logatome, we automatically present only three boundaries to the segmenter: beginning of the diphone, border between the two phonemes, end of diphone.

In a second pass a group of ten to twelve trained phoneticians manually correct the pre-segmented sentences and logatomes. After that three phoneticians that are consistent to each other corrects the segmentations in a third pass. In a last step all segmentations are reviewed by the team supervisor.

To provide a highly consistent phonetic annotation we use the following rules of annotation:

- the placing of boundaries is primarily based on the auditory judgement.
- the boundaries of segments are always placed at positive zero-crossings of the oscillogram.
- the placement of the boundaries should be controlled by sonagram and oscillogram.
- In transitions in which both of two adjacent phonemes can be heard, the boundary is placed in the middle of this transition (50% rule).
- voiced (periodic) elements start with the first clearly identifiable glottal pulse.
- the boundaries of segments with low intensity (e.g. /h/, aspiration) are set where the signal can be clearly distinguished from the background noise. Noises of breathing – if clearly recognised – have to be cut off from the friction or aspiration.
- a lip smack in a sentence is indicated as a /§/. A lip smack within the target diphone of a logatome is not accepted. The logatome has to be recorded again.

Every logatome of every speaker will be segmented manually. Here accuracy is of great importance and we believe that the quality of manually segmented logatomes is still significantly better than phonemes segmented by MAUS.

Since the segmentation is very time-consuming (about six to ten hours for labelling and segmentation for one minute of speech) we currently consider an alternative way for segmenting the unit selection corpus: The quality of the MAUS version can be considerably improved by using an iterative technique: the segmented and labelled speech data of the target speaker are used to re-estimate the acoustical models and the MAUS procedure is applied again to the speech data using those speaker-dependent models ([10]). Considerable improvement can be achieved for a data set larger than one hour of a single speaker. In the BITS context this equals to roughly 300 sentences. Since the unit selection set will comprise about 1700 sentences there might be a good chance to achieve qualitatively sufficient segmentations using MAUS. Whether a manual re-validation of these will be necessary is still undecided yet.

The prosodic annotation of the unit selection corpus is still undecided. The problem here is that the type of annotation considerably influences the type of unit selection algorithm that may be applied together with the corpus. On the other hand the corpus should be general in the sense that as many users as possible may make use of the provided annotations. We would very much welcome any input about that particular issue.

7. Quality Control

After each recording session random samples are checked by a technical supervisor to detect technical noise or other errors. To control the synchrony of the two recorded microphone channels the cross correlation between the signal of the close talk microphone and the signal of the condenser microphone is checked on a regular basis. The delay is basically caused by the distance from speaker mouth to condenser microphone. With 60 cm distance and an assumed acoustic velocity of 330 m/s the theoretical delay is 1.8ms. In practice values between 1.1ms and 2.1ms have been found.

Every sentence and every logatome is saved in an individual WAV file and an individual session number is assigned to the file (e.g. US10031034). The first four digits represent the speaker (1000 – 1003), the last four digits represent the number of the prompt. The prefix codes US and LG stand for unit selection sentences (US = Unit Selection) or logatomes (LG) respectively. The program MAUS automatically creates a TextGrid file with the pre-segmentation for every channel 1 file. The files are then edited manually using the software "Praat" ([6]).

In every pass of the annotation recordings may be declared to be faulty and automatically passed back to the recording group. At the moment roughly 10% of all recordings are repeated due to errors in the signal quality or due to the spoken content.

When the segmentation of a file is finished the segmenter has the possibility to give comments about particular pronunciation variants, tolerated noises, etc. which are inserted into the corresponding SAM label file.

8. Distribution/Availability

Being a publicly funded speech resource the final BITS Speech Synthesis Corpus will be available via BAS [7] or ELDA [8] probably end of 2005. Although there will be no explicit royalties on this corpus the basic distribution fees of BAS will nevertheless apply to this resource as to all other BAS resources to ensure the further maintenance of the resource and long-term availability to the scientific community.

References

[1] http://tcts.fpms.ac.be/synthesis/mbrola.html
[2] http://www.bas.uni-muenchen.de/Forschung/BITS
[3] A. Schweitzer et al. (2003): Restricted Unlimited Domain Synthesis. In: Proceedings of Eurospeech 2003 (Geneva), vol. 2, pp. 1321 – 1324.
[4] Chr. Draxler, K. Jänsch (2004): SpeechRecorder – a Universal Platform Idependent Multi-Cahnnel Audio Recording Software. Proceedings of the LREC 2004 (Lisbon, Portugal), to appear.
[5] F. Schiel (1999): Automatic Phonetic Transcription of Non-Prompted Speech; in: Proceedings of the ICPhS 1999. San Francisco, August 1999. pp. 607 – 610.
[6] http://www.fon.hum.uva.nl/praat/
[7] http://www.bas.uni-muenchen.de/Bas/
[8] http://www.elda.fr
[9] http://www.bas.uni-muenchen.de/Bas/BasSAMPA
[10] F. Schiel (2004): MAUS Goes Iterative. Proceedings of the LREC 2004 (Lisbon, Portugal), to appear.