SpeechDat-AT: A telephone speech database for Austrian German

Micha Baum
Gregor Erbach
Gernot Kubin

FTW
Telecommunications Research Center Vienna
Maderstr. 1/9, A-1100 Wien, Austria
speech@ftw.at

Abstract

We present two telephone speech databases for Austrian German. The databases contain one thousand calls each, from the fixed and from the mobile telephone network. Speakers were chosen to assure a representative distribution over accent regions, sex, and age groups. The databases are compliant with the guidelines of the Speechdat project. We discuss the characteristics of Austrian German, describe the contents of the databases, the speaker recruitment methods, transcriptions of the calls, the phonetic lexicon, and quality assurance measures. In the end, we outline our plans to use the database for research on the pronunciation of Austrian German and automatic dialect identification.

1. Characteristics of Austrian German

There are two basic theories about what is "Austrian German". One is, that the German spoken in Austria is the number of dialects within the area of the Republic of Austria, as there are quite a lot in Germany as well. The other is, that German is a language with more than one centre, a so-called pluricentric language (Muhr, 1997). This is not only valid for dialects but also for standard language, as it used e.g. by professional speakers or politicians. In the first case the two speech databases for Austria can be considered an addition to the German databases in order to receive samples of all dialects and accents within the German-speaking area[1]. In the latter case the purpose of the Austrian databases is to receive one variety of the German language, whose character differs in general from the German spoken in Germany. Speakers tend to imitate such standard language, they hear in the public and they are spoken to e.g. in a dialogue system.

Apart from the acoustic deviations there are lexical differences in the German speaking countries for which speech data are needed, e.g. *Jänner* vs. *Januar* (January) or Austrian city names.

2. Contents of the Database

SpeechDat-AT is a telephone speech database collection. This document describes the full 1000 speaker databases of Austrian SpeechDat-AT recordings from the fixed network (FDB) and the mobile network (MDB).

2.1. Items

The final specification for the recordings is as follows:

48 items are mandatory for the DBs of SpeechDat-AT, and 8 are optional. The optional items are included in the database because they provide speech useful for teleservices: responses to speaker gender questions, telephone networks, birthdate utterances (whose format may differ considerably from other date utterances), today's date (which is assumed to be the most important date information in teleservices), the names of all Austrian federal states, environment of call, native language (mostly "Österreichisches Deutsch") and educational level. On the other hand these items contain speaker-specific material, which is mandatory or at least optional for the label files.

Some of the items are read, others contain spontaneous speech.

2.1.1. DIGITS

I1/I2	isolated digit
B1	sequence of 10 isolated digits
C1	prompt sheet number = 7 digits
C2	telephone number (9-11 digits)
C3	credit card number (15-16 digits)
C4	PIN code (6 digits)
N1	natural number
M1/M2	money amount

2.1.2. YES/ NO QUESTIONS

Q1	spontaneous predominantly yes
Q2	spontaneous predominantly no

2.1.3. DATES

D1	spontaneous date
D2	prompted text form
D3	relative and general date form

2.1.4. TIMES

T1	time of day, spontaneous
T2	prompted, mixed analogue/digital

2.1.5. APPLICATION WORDS

A1-6	keywords/keyphrases

2.1.6. WORD SPOTTING PHRASE

E1	using embedded application words

2.1.7. DIRECTORY ASSISTANCE

O1	names spontaneous (forename)

[1] There is another speech database for Swiss German.

O2	city of birth/growing up (spontaneous)
O7	set of 150 full names
O3/O4	most frequent cities
O5/O6	most frequent companies, agencies

2.1.8. SPELLINGS

L1	spontaneous spelling, e.g. forename
L2	directory city name
L3	real/artificial word

2.1.9. PHONETICALLY RICH WORDS

W1-4

The phonetically rich words were selected from the original lexicon in such a way that words with rare phonemes had a higher priority than more normal words[2].

2.1.10. PHONETICALLY RICH SENTENCES

S1-9, Z0-1

The sentences are taken from newspapers (Süddeutsche Zeitung, Frankfurter Rundschau, Der Standard) or homepages (www.orf.at). The following criteria were used:

- ***Not too short***

In order to obtain a sufficient number of phonemes in each sentence, the minimal number of words per sentence is three.

- ***Not too long***

It was considered important to keep the sentences readable, that means, make callers pronounce the prompted sentences. Therefore, as many sentences as possible consists of less than 14 words. In a few cases there were not enough short sentences with some rare phonemes, so they have more than 14 words.

- ***No punctuation***

Since punctuation is a sign of complex sentence structure, we selected sentences without punctuation as far as possible.

On each prompt sheet there are 12 different phonetically rich sentences (items S0-S9, Z0-Z1). In each of these corpora one rare phoneme was selected. So every phoneme, we defined as native, occurred at least once within the phonetically rich sentences, which was one of the criteria we had to fulfil. Here is a list of the 12 rare phonemes in Sam-Pa notation and their corresponding items:

S0	tS	S6	Y
S1	pf	S7	y:
S2	2:	S8	j
S3	E:	S9	x
S4	9	Z0	u:
S5	OY	Z1	aU

For each of these items 480 different sentences were selected. That makes 5760 different sentences for the whole database.

The phonemes /a~/, /o~/, /Z/, and /dZ/ are used in the lexicon and are counted, but are not considered mandatory. They may occur but they need not.

2.2. Speaker-specific material

The speakers were asked a number of questions about themselves, such as their gender, accent region, phone network etc. These questions serve a dual purpose: on the one hand, these are words which are likely to be useful in teleservices. On the other hand, they provide demographic information about the speaker, even if they do not send in their questionnaire.

Y1	speaker gender question
Y2	call from fixed network or mobile network
Y4	speaker region question
Y5	today's date
Y6	environment of call
Y7	native language
Y9	educational level

2.3. Speaker Demographic Information

2.3.1. ACCENT REGIONS

The federal states of Austria were chosen as accent regions. This information can be provided easily by the speaker, and 9 regions can be administrated relatively easily and also coincide with accent boundaries (e.g. the river Enns, or the Arlberg mountain).

Figure 1 and figure 2 show the accent regions and the number of callers from each of them as well as the targeted number of callers[3].

2.3.2. SEX/AGE

Table 2 and table 3 show the distribution over sex and age groups.

Age Group	Gender	Count
0–15	F	9
0–15	M	6
16-30	F	212
16-30	M	232
31-45	F	137
31-45	M	191
46-60	F	83
46-60	M	101
61-99	F	15
61-99	M	14
total	F	456
total	M	545

Table1: Sex and age groups (FDB)

[2] Every phoneme has to occur at least 100 times within the phonetically rich words.

[3] The MDB was not finished yet at the deadline of this paper.

Age Group	Gender	Count
0–15	F	11
0–15	M	12
16-30	F	205
16-30	M	349
31-45	F	87
31-45	M	179
46-60	F	48
46-60	M	59
61-99	F	5
61-99	M	3
total	F	356
total	M	602

Table2: Sex and age groups (MDB)

2.3.3. ENVIRONMENTS

For the MDB the calls were made from following environments, with the goal, to reach at least 200 from each of them:

home/office
public place (i.e restaurant)
moving vehicle
pedestrian by road side

For the FDB the only goal, concerning environments, was to reach 20 calls from a public place. The FDB includes 121 of them.

2.4. Label files

Each speech file is associated with one label file which contains information about the recording and the speaker, as well as the prompted text and the orthographic transcription.

The labels files use US-7bit ASCII SAM labels. The meanings of the labels are as defined in SD1.3.1 (Senia 1997). A sample file is given below.

```
LHD: SAM, 5.1
DBN:
SpeechDat_East_Austrian_FIXED_Network
VOL: FIXED1AT_01
SES: 1003
DIR: \FIXED1AT\BLOCK10\SES1003
SRC: A11003A1.ATA
CCD: A1
CRP:
BEG: 0
END: 26471
REP: Connect Austria, Vienna
RED: 29/Oct/1999
RET: 16:21:56
SAM: 8000
SNB: 1
SSB: 8
QNT: A-LAW
SCD: 001003
SEX: M
AGE: 24
ACC: OOE
REG: Wien
ENV: OFFICE
NET: FIXED
PHM: NORMAL
LBD:
LBR: 0,26471,,,,löschen
LBO: 0,,26471, löschen
SHT: 200-0202
EDU: HIGH SCHOOL
NLN: DE-AT
ASS: OK
ELF:
```

3. Transcription

The transcription software used (WWWTranscribe) was developed by the Department of Phonetics at the University of Munich, Germany (Draxler 1997). It is based on the WWW and thus allows transcriptions on any machine with a modern WWW browser. The signal is presented auditively only and can be output as often as requested. For ease of editing, some special conversion buttons have been implemented, e.g. digit to number string, spelling to letter names (A ➠ Anton, B ➠ Berta) etc.

The character set used for the transcription is ISO 8859-1 (ISO Latin).

Transcriptions make use of the standard German spelling before the recent spelling reform. We follow the German rules for capitalization, i.e. nouns, names and nominalized verbs are written with capital letters, everything else with lower case letters.

Punctuation marks are removed from the transcription.

The following symbols are used to denote word truncations, mispronunciations, non-understandable speech, non-speech acoustic events; other symbols:

~abc or abc~	word truncations	at signal begin or end only
*abc	mispronunciations	
**	non-understandable speech	separated by blank from rest of text
[fil]	filled pause	at correct location between words
[spk]	speaker noise	at correct location between words
[int]	intermittent noise	at correct location between words
[sta]	stationary noise	at correct location between words

Table 3 - Noise and mispronunciation markers

4. Recording site and platform

The recording platform are two 64 MB Pentium III 450 PC, running Windows NT. Both machines are equipped with two 8.6 GB internal mirrored hard disks and a Fritz 32PCI primary rate ISDN interface board. As an additional backup tool an external DAT drive was used.

The software for recording (ADA = Automatic Database Acquisition using a CAPI-based ISDN call server) was developed at the "Universitat Politècnica de Catalunya" in Barcelona. The software allows a maximum of 4 parallel recordings; if more than 4 calls are received in parallel, all calls in excess of 4 receive a busy signal and are not answered.

Each call is given a unique ID-number, and no attempt is made to identify a caller. Hence, if a call is aborted, it will be incomplete on the hard disk.

5. Speaker Recruitment

Speaker recruitment was performed by a variety of methods:

Promptsheets could be requested via the Internet (http://speech.ftw.at/) or by a phone call to a free number. Approximately 250 out of 2400 callers requested promptsheets by telephone, the others obtained the promptsheets from our web site.

Method	Responses
Media (print and broadcast) coverage by means of press releases	approx. 350
Links which were generously provided by a number of websites (ONE.at, comparatel.at, handyinfo.at, handy.at)	179
Postings in appropriate newsgroups (at.telekomm.* etc.)	263
E-mails to target groups (participants in previous data collections, linguists)	approx. 350
Information of staff of partner companies	approx. 500
Snowball system: participants in the speech database were rewarded for recruiting further speakers	1577
Individual contacts by a market researcher	100

Table 4: Speaker recruitment

5.1. Rewards

For each call, the caller received five lottery tickets (value 10 ATS = 0.73 Euro each) or speech recognition software (Philips SpeechMania 98, Austrian version).

The reward for recruiting a caller was initially one lottery ticket per call. In later stages this was increased to 3 and 5 lottery tickets, in accordance with the difficulty of finding a person from one of the missing groups.

With each reward that was sent out, we also sent a letter which provided information about the types of callers still missing from the database and a request to recruit more callers.

5.2. Problems encountered

Since most of our speaker recruitment was done over the Internet, we encountered the problem that the following groups, which constitute the majority of the Austrian Internet population, were over-represented:

- Male persons
- Age 16 – 45 (especially below 30 years)
- Callers from Vienna

This problem was overcome by closing the data collection to groups from which we had a sufficient number of calls, and at the same time increasing the rewards for recruiting a speaker from a missing group.

In order to obtain calls from underrepresented accent areas, we contacted people from these areas who had already participated in the data collection, and offered them rewards for recruiting more speakers from the same region. This method was successful in filling gaps in the database.

6. Quality Assurance

We took measures to ensure that the transcriptions of the recorded calls were error-free and consistent. Transcribers were chosen that had a background in phonetics and/or professional experience with handling German texts (e.g. translators).

6.1. Continuous QA

Clear guidelines for transcription were established and handed out to every transcriber. Each new transcriber had to read the guidelines and transcribe one session. Then the transcriptions were discussed with the transcriber and the errors corrected.

The transcriptions were checked at regular intervals, and transcribers alerted to their errors. Transcriptions were also checked against the lexicon to find spelling errors. Usage of special signs was checked separately.

Five percent of the calls were checked after collection of the database to obtain an estimate of the error rate.

6.2. Validation

Validation of the database is carried out by the Speech Processing Expertise Center (SPEX) in Nijmegen. In order to avoid errors in the production of the database, a preliminary versions of the fixed and mobile database (10 calls each) were delivered to SPEX for pre-validation.

The entire database will be formally validated by SPEX at the end of the project.

7. Phonetic Lexicon

The SpeechDat-AT lexicon is mainly provided by Philips Vienna. It covers all texts on the prompt sheets. The transcription of the speaker’s utterances were not completely covered by this lexicon. These missing words had to be transcribed by hand (primarily names and specific Austrian vocabulary). We tried to stick as close as possible to the pronunciation rules of the "Ausspracheduden" (Drosdowski et al. 1990).

The SAM-PA symbols used are those of the German SAM-PA inventory with two additions: the ~ was used to indicate nasalization of /O/ and /a/ which occur in words of French origin (e.g. „Restaurant“). Not all diphthongs allowed in the SAM-PA table were used. The pure free vowels were transcribed with a colon in any case.

The lexicon entries are case sensitive and follow the German rules for case, i.e. nouns and names are written with the first letter a capital letter, all other words are written in lower case. Note that the polite form of personal pronomina, e.g. „Du“, „Sie“ are written in lower case too; nominalized verbs, e.g. „das Laufen“ are treated as nouns and hence are written with a capital first letter.

A lexicon entry may contain hypens, apostrophes, or slashes. Idiomatic expressions are split into their word constituents.

Only one spelling of a word is allowed. This spelling is determined by the Duden or the source from which the original word was taken (e.g. newspaper text for words not in the Duden). Alternative spellings are considered to be different words and hence they are included in the dictionary as entries of their own. This is especially true for foreign and loan words that have both a German and the original spelling, e.g. „Telefon“, „Telephone“.

Note that the orthograpy reform of German is not applied to the lexicon.

8. Future Research

The speech databases will be used for further research in the project.

8.1. Speech Dialog System

In cooperation with the Austrian Post Office, we will build a speech dialog system which provides information about Austrian postal rates. We use the telephony speech dialog system Philips SpeechMania as the basis for this development. The acoustic models of SpeechMania will be trained for Austrian German making use of the SpeechDat-AT database. The system will become operational in September.

We plan a comparative evaluation of the speech recognition performance with the German and Austrian acoustic models.

8.2. Automatic identification of dialects

We will use the speech database as the basis for research concerning the automatic identification of dialects by making use of intonation differences. For this purpose, we also use a database with German speakers. While intonation has not proven to be the best method for language identification (Muthusamy et al. 1994), we expect it to be useful in the identification of dialects, which do not differ significantly in their phoneme inventory. This research is carried out in cooperation with the University of Music and Performing Arts in Graz.

8.3. Pronunciation models for Austrian German

The database includes a pronunciation lexicon based on German standard pronunciation because no pronunciation lexicon for Austrian German exists. However, we believe that appropriate modeling of Austrian pronunciation would help improve speech recognition performance. We will work on pronunciation rules for Austrian German, with the intention of capturing regular differences between Austrian German and standard German, and the semi-automatic construction of a pronunciation lexicon for Austrian German and its variants.

9. Intellectual Property Rights

Usage of the speech database for commercial purposes is restricted to the project participants. Parts of the database may be made available for scientific research on a case-by-case basis.

10. Acknowledgements

This work was supported within the competence center program ***Kplus***, and by the companies Alcatel, Connect Austria, Kapsch, Nokia, Philips Speech Processing, and Siemens.

11. References

Draxler, Christoph (1997). WWWTranscribe: A modular transcription system based on the WWW. Proceedings of Eurospeech, Rhodos.

Drosdowski, Günther; Müller, Wolfgang; Scholze-Stubenrecht Werner; Wermke Matthias (1990). Das Ausssprachewörterbuch. Dudenverlag Mannheim

Muhr, Rudolf (1997). Norm und Sprachvariation im Deutschen. Das Konzept "Deutsch als plurizentische Sprache und seine Auswirkungen auf Sprachbeschreibung und Sprachunterricht DAF. In *Germanistische Linguistik* Gerhard Helbig, Marburg an der Lahn

Muthusamy, Y.K.; Barnard, E.; Cole, R.E.; (1994). Reviewing Automatic Language Identification. *IEEE Signal Processing Magazine*, Oct. 1994, 33-40.

Senia Francesco (1997). Specification of Speech Database Interchange Format. LE2-4001 SpeechDat Deliverable 1.3.1

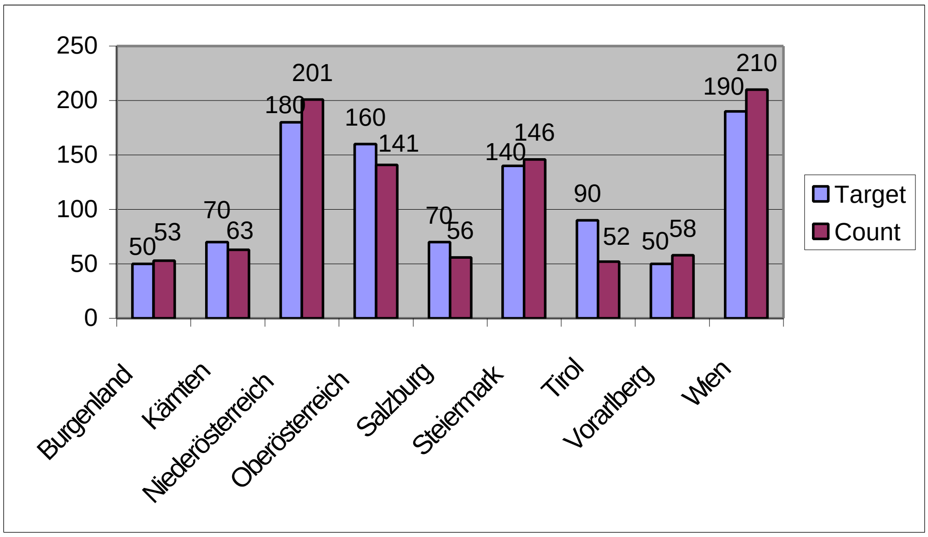

Figure1: Accent Regions (FDB)

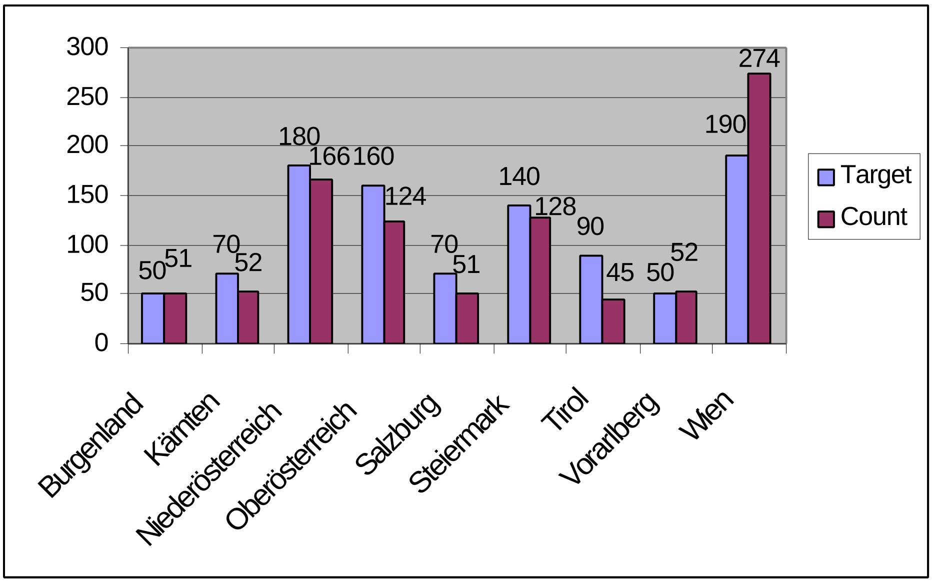

Figure2: Accent Regions (MDB)