A CROSS-DATABASE COMPARISON OF TWO LARGE GERMAN SPEECH DATABASES

Christoph Draxler[1], Stefan Kleiner[2]

[1]Institute of Phonetics and Speech Processing, Ludwig Maximilian University, Munich

[2]Institute of German Language, Mannheim
draxler@phonetik.uni-muenchen.de, kleiner@ids-mannheim.de

ABSTRACT

Ph@ttSessionz and Deutsch heute are two large German speech databases. They were created for different purposes: Ph@ttSessionz to test Internet-based recordings and to adapt speech recognizers to the voices of adolescent speakers, Deutsch heute to document regional variation of German. The databases differ in their recording technique, the selection of recording locations and speakers, elicitation mode, and data processing.

In this paper, we outline how the recordings were performed, how the data was processed and annotated, and how the two databases were imported into a single relational database system. We present acoustical measurements on the digit items of both databases. Our results confirm that the elicitation technique affects the speech produced, that f0 is quite comparable despite different recording procedures, and that large speech technology databases with suitable metadata may well be used for the analysis of regional variation of speech.

Keywords: speech database, acoustic analysis, regional variation, speech technology, metadata

1. INTRODUCTION

Speech databases are created for many different purposes, and this determines the design and contents of such databases. For example, speech technology development requires speech databases large in terms of phonetic variation and application specific vocabulary, whereas language variation and sound change require databases with precise geographical and temporal information.

There are a number of problems related to speech databases. First, only for a few selected languages such databases exist in sufficient numbers and quality; for most languages, few or no such databases exist. Second, the diversity of application areas and languages plus the general focus on creating databases mainly for one's own purpose leads to very heterogeneous database formats, making re-use or a comparative analysis difficult. Finally, there are no (or very few) software tools that are both tailored to the specific needs of speech researchers and capable of dealing with very large databases. Emu was one of the first systems to allow queries over an entire database of speech data, but its query language is restricted and performance sharply degrades with database size [5, 4], EXAKT is a query system for EXMARaLDA speech databases, but focuses on the annotation [16], and COREX was developed specifically for the CGN database [12] – to name a few.

In this paper we present a simple unified data model for speech databases and implement it using a relational database system (cf. [1] for an object-oriented approach). We import two quite different speech databases and formulate compararative queries to analyse the effects of recording equipment, elicitation modes, and data processing on selected acoustic measures.

2. PH@TTSESSIONZ

The Ph@ttSessionz (PHAT) speech database was designed a) to demonstrate the feasibility of large-scale Internet-based speech recordings, and b) to provide training data for speech recognizers for voices of adolescent speakers [7]. For compatibility reasons, the speech material follows the design of the SpeechDat databases, i.e. it contains digits, numbers, time and date expressions, spellings of person, company and geographic names, phonetically rich words and sentences [13, 8]. Furthermore, it contains a number of non-scripted utterances.

2.1. Recording locations

Schools all over Germany were asked to participate in an innovative science project for technology development, and they received a lump sum of 300€ for 30 recorded speakers. To participate, a school had to name a recording supervisor, in general a class teacher or pupils from higher classes. This person was then responsible for recruiting speakers.

In total, schools in 46 different cities contributed recordings to PHAT, with a total of 1019 recording sessions. On average, 22.15 sessions were recorded in each city, the median is 24 sessions.

2.2. Equipment

The recording equipment consisted of an M-Audio Mobile Pre USB microphone amplifier and A/D converter, a Beyerdynamic opus 54 close-talk microphone and an Audio Technica 3031 desktop microphone. The signal quality is 22.05 kHz sampling rate with 16 bit linear quantisation and stereo.

2.3. Software

The recordings were made via the Internet using a standard browser. To start a session, the speaker entered her or his demographic data via a form. Then, the recording application progressed through a recording script. Each utterance was written to a separate audio file [6].

Per session, a total of 131 items were recorded. On average, a session lasted for approx. 17 minutes, producing 5 min of speech signal, resulting in a total of 83 hours of transcribed speech signal.

2.4. Postprocessing

All audio files were transferred to the server during the recordings. The flac audio files were expanded to a non-compressed format, split into two channels and saved in WAV format. A listening test determined whether the channels were correctly labelled.

2.5. Annotation

PHAT was transcribed orthographically following the SpeechDat transcription guidelines. For this, a dedicated web-based transcription tool was used, so that transcribers could work in the office or from home. The transcription tool implements a simple lexical analyzer to ensure that only syntactically correct transcriptions were entered into the database. Transcribers listened only to the close-talk channel recordings.

The orthographic transcriptions were then automatically segmented and labelled on three annotation levels by the MAUS system [14].

3. DEUTSCH HEUTE

The speech database Deutsch heute (DEHE, *German today*) was designed by a linguistic project group at the Institute for the German Language (IDS) to gather information on regional variation primarily in pronunciation, secondarily also in lexical, morphological and syntactical variation, of Standard German in the whole area, where German has (co-)official status. Since 2011, main results of the corpus analysis have been continuously published in a wiki-based linguistic atlas project (AADG, http://prowiki.ids-mannheim.de/bin/view/AADG). Other speech databases containing regional variants of German hosted by the IDS are e.g. the contemporary FOLK-Corpus (currently 70 h of dia- or multilogues) or the 1950s Zwirner-Corpus, which comprises more than 5000 10-minute recordings (mainly of narratives) of regional varieties from all over the pre-war German speaking area. Elsewhere especially the Deutscher Sprachatlas at Marburg has a large repository of speech data of regional varieties of German, parts of which can be accessed via the REDE-website (http://www.regionalsprache.de/).

The corpus design of DEHE follows and expands the one devised by [10] for his linguistic atlas of the former FRG. It consists of read and spontaneous speech (of app. 45min each per participant). The read speech part contains three texts: 1. The North Wind and the Sun read at normal and fast speaking rate, 2. an 800-word text, 3. a 500-word text from a popular scientific journal, 4. a 1000-word word list, 5. a picture naming task and 6. a translation from English into German.

The spontaneous speech part contains a 30 min socio-biographic interview with a linguist and a 15 min map task experiment (cf. [2]) between two participants. Relevant biographic data was collected in a questionnaire.

3.1. Participants and recording locations

In order to document language change in apparent time, two age-groups were recorded. The main group of 671 participants consisted of secondary school students (age 16-20), the secondary group of 158 participants with secondary school education (age 50-60). The participants, which had to be born and raised locally (which also applied to at least one of their parents), were recruited by teachers and employees at cooperating local schools and adult education centres.

Recordings took place in 194 places all over the German-speaking area (Germany 146, Austria 25, Switzerland 13, Italy (South Tyrol) 3, Luxemburg 2, East Belgium 2, Liechtenstein 1). The recordings were made locally by linguists in a quiet room on the school or education centre premises. At each place usually four students were recorded.

3.2. Equipment

For the recordings of the younger speakers, Marantz PMD 671 solid-state recorders with Sennheiser HSP 4 condenser cardioid neckband microphones were used.

The signal was recorded at 44.1 kHz, 16 bit and stored in wav-file format. In the read speech parts only one channel was recorded, in the interviews and map tasks each of the two speakers were recorded on a separate audio channel.

3.3. Annotation

All parts of the corpus were transcribed orthographically. The spontaneous speech parts and read texts were segmented and aligned on the phrasal level, the word list, translation and picture naming tasks were segmented and aligned on the word level. For this task Praat [3] was used. For all prompted speech parts a canonical phonetic transcription was generated by using the MAUS aligner [14].

4. DATA BASE DESIGN

Ph@attSessionz is available via the online repository at the Bavarian Archive for Speech Signals (BAS, http://clarin.phonetik.uni-muenchen.de), DEHE is available for research purposes upon request to the IDS.

Both speech databases come in different formats: PHAT is formatted according to the BAS standard distribution format, i.e. one directory for every recording session, plus metadata in CMDI format. DEHE is distributed on DVDs, one directory per recording location, and one long WAV-file per task, with matching Praat TextGrid files.

4.1. Relational databse

Both speech databases were imported into a PostgreSQL relational database system (http://www.postgresql.org). This not only allows queries directly on the database using the standard SQL language, it also supports access by most programming languages, web servers, and statistics software.

The data model is shown in figure 1. It consists of 8 interconnected tables. Tables *geolocation* and *segment* contain hierarchical data: a country has federal states which in turn have cities, and transcripts consist of words in sequence which in turn contain time-aligned phonetic segments. These hierarchies are represented using links within tables.

Table 1 displays the contents of the two databases for recordings in German cities and speakers between 16 and 21.

Figure 1: Relational datamodel of the database

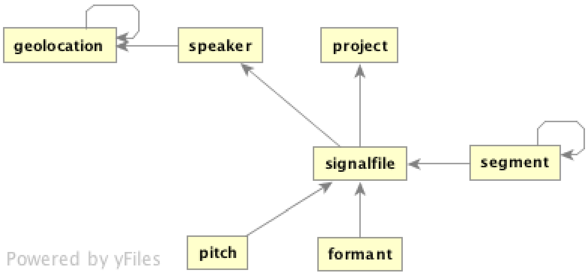


Table 1: Database contents for speakers aged 16-21 recorded in Germany

Database	speakers (m/f)	federal states	cities (m/f)
DEHE	478 (229/249)	16	118/121
PHAT	520 (253/267)	14	41/37

The two federal states not covered by PHAT are the city states of Bremen and Hamburg.

5. SELECTION OF DATA

For the comparison of the two speech databases, a subset containing only digits was selected. Digits are interesting because they can be presented in a non-orthographic numerical format, occur in every sociolect or dialect region, and they are, with the exception of the digit 7 ('sieben'), monosyllabic. The selected digit items are given in Table 2.

DEHE does not contain the digit 0. Only 7 (1.46%) speakers produced *zwo* for the digit 2. In PHAT speakers chose to use *zwei* 490 (95.52%) times vs. 23 (4.48%) times for *zwo*.

6. ACOUSTIC MEASUREMENTS

The following acoustic measures were computed using formant and pitch tables pre-calculated by Praat with the default settings. The evaluation of the SQL queries took between 1 and 18 seconds on a standard laptop computer.

6.1. Word and phoneme durations

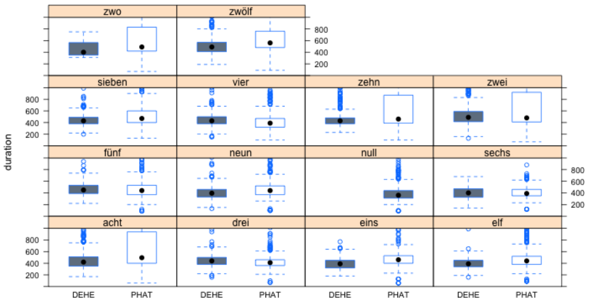
In PHAT, digits were elicited one by one, whereas in DEHE they were spoken as a word list, using the prompt 'ZAHLEN von 1-25' (*numbers from 1-25*). word lists are known to affect the speech produced: e.g. by rhythmic grouping, or pre-final lengthening.

Table 2: German digit pronunciation in SAM-PA and item count per speech database

digit	word	phonemes	PHAT	DEHE
0	null	/n U l/	516	
1	eins	/aI n s/	520	478
2	zwei	/ts v aI/	493	471
	zwo	/ts v o:/	506	7
3	drei	/d R aI/	519	478
4	vier	/v i:6/	521	478
5	fünf	/f Y n f/	518	478
6	sechs	/z E k s/	518	478
7	sieben	/z i: b @ n/	511	478
8	acht	/? a x t/	512	478
9	neun	/n OY n/	520	479
10	zehn	/ts e: n/	515	478
11	elf	/? E l f/	518	478
12	zwölf	/ts v 9 l f/	514	478

To analyse whether the elicitation also affects speaking rate, we compare word and phoneme duration. When producing words in a word list, speakers typically increase their speech rate so as to finish the recordings quickly.

Figure 2: Word duration per database



The durations for most words are very similar, but for words with an initial or final plosive the difference is marked (cf. figure 2). This is due to impossibly long segment durations for the plosive /t/, and they are an artefact of the automatic segmentation: MAUS cannot reliably determine the exact start of the plosive closure in initial, or the exact end of the aspiration in final position.

This effect does not show up in DEHE because the affected phonemes do not occur at the beginning or end of the digit sequence.

Excluding /t/, the average duration of phonemes in DEHE is 117ms vs. 137ms in PHAT. This is a difference of 14.6%, which can be attributed to the word list elicitation mode.

6.2. f0 per age

As expected, there was no significant f0 vs. age difference between the two databases.

6.3. Vowel formants

Vowel formants are often used to analyse regional variation in speech, and although the phoneme inventory of digits is limited, it features interesting phenomena. From the dialectological literature it is known that speakers from Bavaria may have a further back /a/, i.e. lower f2), and speakers from Saxony and Thuringia a more centralised initial /E/, i.e. higher f1 (cf. [11, 9, 15]).

Table 3: Vowel formants for /a/ and /E/

Database	word	vowel	state	f1	f2
DEHE	acht	a	BAV	760	**1316**
			other	796	1408
PHAT		a	BAV	767	**1278**
			other	815	1388
DEHE	elf	E	EAST	**603**	1922
			other	586	1922
	sechs		EAST	516	2002
			other	522	1964
PHAT	elf	E	EAST	**605**	1988
			other	561	1948
	sechs		EAST	513	1939
			other	510	1963

Table 3 shows f1 and f2 for /a/ and /E/. In fact, /a/ is futher back in Bavaria (BAV), and initial /E/ is more centralisd in Thuringia and Saxony (EAST).

7. CONCLUSIONS

The speech databases PHAT and DEHE were collected for different purposes and using different recording equipment and techniques. Elicitation via word lists vs. individual prompts affects phoneme and thus also word durations: for word lists, especially for highly automated tasks such as counting from 1 to 25, speech rate increases considerably. Automatic segmentation of phonemes is more error prone for individual items than for word lists, especially for initial and final plosives. The differences in technical equipment, recording procedure and elicitation technique do not show marked effects on the acoustic measures f0, f1, and f2, and this suggests that speech technology databases may well be used for research in regional variation – and perhaps also vice-versa.

8. REFERENCES

[1] Altosaar, T., Millar, B., Vainio, M. 1999. Object-oriented models for representing speech: A comparison using ANDOSL data. *Proc. Eurospeech.*

[2] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R. 1991. The HCRC map task corpus. *Language and Speech* 4(34), 351–366.

[3] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341–345.

[4] Cassidy, S. 1999. Compiling multi-tiered speech databases into the relational data model: experiments with the emu system. *Proc. Eurospeech* Budapest. 2239–2242.

[5] Cassidy, S., Harrington, J. 2001. Multi-level annotation in the emu speech database management system. *Speech Communication* (33), 61–77.

[6] Draxler, C. 2006. Exploring the Unknown – Collecting 1000 speakers over the Internet for the Ph@ttSessionz Database of Adolescent Speakers. *Proc. Interspeech* Pittsburgh, PA. pp. 173–176.

[7] Draxler, C., Steffen, A. 2005. Ph@ttSessionz: Recording 1000 adolescent speakers in schools in Germany. *Proc. Interspeech* Lisbon. 1597–1600.

[8] Höge, H., Draxler, C., van den Heuvel, H., Johansen, F., Sanders, E., Tropf, H. 1999. SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line. *Proc. Eurospeech* Budapest. 2699–2702.

[9] Kehrein, R. 2012. Regionalsprachliche Spektren im Raum – zur linguistischen Struktur der Vertikale. *ZDL Beihefte* (152).

[10] König, W. 12. Auflage, 1998. *dtv-Atlas Deutsche Sprache*. Deutscher Taschenbuch Verlag.

[11] König, W., Renn, M. 2006. *Kleiner Bayerischer Sprachatlas*. Deutscher Taschenbuch Verlag.

[12] Oostdijk, N., Broeder, D. 2003. The spoken dutch corpus and its exploitation environment. *Proc. 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)* Budapest. pp. 93–100.

[13] Pearce, D. 1995. Specification of short/mid-term databases. Technical report LRE-63314 SpeechDat(M) Report D1.3.1 & D1.3.2.

[14] Schiel, F. 2004. MAUS goes iterative. *Proc. LREC* Lisbon, Portugal. 1015–1018.

[15] Schimunski, V. 2010. *Deutsche Mundartkunde: vergleichende Laut- und Formenlehre der deutschen Mundarten.* Frankfurt, Berlin, Berin, Wien et al.: Peter Lang Verlag.

[16] Schmidt, T., Wörner, K. 2012. Introduction. In: *Multilingual Corpora and Multilingual Corpus Analysis* volume 14 of *Hamburg Studies in Multilingualism*. John Benjamins ix – xi.