The German SpeechDat Telephone Speech Corpus
Overview and Experiences

Christoph Draxler

Department of Phonetics and Speech Communication
University of Munich

ABSTRACT
SpeechDat is a European project to collect ISDN quality telephone speech for all major European languages. In the first phase of the project, 1000 speakers were recorded in eight languages, including German. The paper presents the experiences made during the data collection for German, and outlines the specifications for the second phase of the project.

INTRODUCTION

The SpeechDat[1] project is a joint industrial and academic effort to collect Polyphone-like ISDN quality telephone speech corpora for major European languages. These corpora provide a common basis for the development of speech processing applications for the European languages, and for phonetic, linguistic, and telecommunications research.

The project is divided into two phases, SpeechDat(M) from September 94 until February 96, and SpeechDat II from March 96 until March 98. SpeechDat (M) served primarily as a case study for the definition of common standards, the technical feasibility of the data collection, and the creation of a data validation and distribution infrastructure. These goals have been achieved successfully:

- the ELRA (European Language Resource Association) has been established and funding is guaranteed by the CEU for three years, and
- all SpeechDat(M) corpora have been validated and pressed on CD-ROM.

The main objective of SpeechDat II is to extend the data collection to larger speaker populations and to include additional languages.

Language	SpeechDat (M)	SpeechDat II	Language	SpeechDat (M)	SpeechDat II
British English	1000	4000	Welsh		2000
Danish	1000	4000	Belgian French		1000
German	1000	4000	Finnish Swedish		1000
Italian	1000	3000	Flemish		1000
Portuguese	1000	4000	Norwegian		1000
Spanish	1000	4000	Russian		1000
French		5000	Slovenian		1000
Greek		5000	Swiss German		1000
Swedish		5000	Luxemburgish French		500
Finnish		4000	Luxemburgish German		500
Swiss French		2000			

Table 1. Speechdat Languages and Speaker Populations

A similar data collection with 5000 callers has been carried out for Dutch prior to the SpeechDat project (den Os et al.,1995). 1000 callers were recorded for Swiss French at the same time when SpeechDat(M) began; this corpus was later included in SpeechDat(M).

Structure of the paper

The remainder of the paper is structured as follows: section 2 contains the SpeechDat corpus specifications, section 3 outlines the technical setup. Section 4 describes the speaker recruitment and

contains statistical information on the speaker population. Section 5 presents the orthographic transliteration. Section 7 discusses the SpeechDat(M) validation, and section 8 summarizes the experiences and gives an outlook on SpeechDat II.

SPEECHDAT DATABASE SPECIFICATIONS

The German SpeechDat(M) corpus consists of 1000 calls. Each call consists of 44 recorded items, 39 of which were mandatory (for all languages) and 5 were optional. In SpeechDat II, the emphasis has shifted towards a better phoneme and diphoneme coverage, and towards real-world speech data such as geographical and company names; furthermore, the application words and phrases have been revised to match the requirements of telephony applications (Winski, 1996).

SpeechDat		Item Type	Specification
M	II		
1	2	isolated digit items	isolated digit, sequence of 10 isolated digits
3	4	digit/number strings	prompt sheet number, telephone number, credit card number, PIN code
3	1	natural number	
2	1	money amount	currency amount, mixed size and units
3	2	yes/no questions	spontaneous replies
3	3	dates	spontaneous date, e.g. birthdate, prompted text form, relative and general date form
2	2	times	spontaneous and prompted time of day, mixed analogue and digital format
6	3	application keywords/keyphrases	
3	1	word spotting phrase using embedded application words	
1	4	directory assistance names	city of birth/growing up (spontaneous), most frequent cities, most frequent companies or agencies, forename (spontaneous)
	1	proper name	set of 150 SDB full names
3	3	spellings	spelling (spontaneous), directory city name, real/ artificial word
0	4	isolated words	
9	9	phonetically rich sentences	
5	11	partner specific material	speaker gender, fuzzy question, birthdate, speaker region, today's date, form task, spontaneous speech, good-bye phrase
44	51	TOTAL	

Table 2. SpeechDat Item List

The exact vocabulary to be recorded was specified separately for each language. In SpeechDat(M), the most restrictive constraint for the corpora was that every prompt sheet must contain each phoneme of the language at least twice (except for very rare phonemes). In most cases this was achieved by selecting the phonetically rich sentences from a large text corpus, e.g. newspaper text.

As a general guideline, the duration of an interview should not exceed 10 minutes. For German, the total duration of an interview was approx. 8 minutes. In a complete interview, approx. 4 minutes of speech were recorded.

Prompt sheets

The SpeechDat(M) prompt sheets consisted of a folded A3 sheet. The first page contained a short motivation and instruction text, and a data form to be filled in and returned to the university. The second and third page contained the prompt sheet itself.

The prompt sheet had a table layout with numbers preceding groups of items. All text spoken by the speech server was printed on the sheet to allow the caller to keep trace of the progress; this text consisted of instructions, feedback messages, and requests for spontaneous speech. The text to be read by the caller was printed in bold font and a bullet was used to represent the system beep.

Bitte lesen Sie Punkt 7 vor.	
7.	• **Aktivieren Sie die Nachricht.** • **Löschen Sie die Umleitung.** • **zurück zur Ansage.**
Vielen Dank. Sie haben die Hälfte des Fragebogens schon hinter sich gebracht.	

Figure 1. Prompt sheet detail

The items for the prompt sheets were held in a relational database management system (RDBMS). For the generation of the prompt sheets, the item texts for each item type were chosen randomly from the RDBMS except for the phonetically rich sentences. These sentences were generated by extracting from the newspaper corpus sets of sentences which met the minimum phoneme number constraint and which were easy to read (i.e. were no longer than 12 words). From the approx. 11.000 sentences of the newspaper text (economics section of a large German newspaper), 112 disjoint sets of 9 sentences each were generated, leading to 112 different prompt sheets. For 1000 recordings, each prompt sheet was used more than once.

In the final validation, the layout of the prompt sheet was criticized because the items were arranged in groups. This could lead to improper pronunciations, e.g. lowering one's tone near the end of a list or fatigue effects. In SpeechDat II the items are distributed over the prompt sheet.

TECHNICAL SETUP

The German SpeechDat(M) recordings were carried out at the Department of Phonetics and Speech Communication at the University of Munich in close collaboration with Siemens AG, Munich. The recording equipment was a 66 MHz 486-DX2 PC running SCO Unix, with 32 MB main memory, an internal 1GB and an external 4 GB hard disk.

The PC was equipped with an NMS (Natural Micro Systems) AG-30 audio board with 6 TMS 320C51 signal processors and an Aculab ISDN board connected to a primary rate (30 channel) ISDN interface. The ISDN installation was not without problems: an improper grounding in the distribution closet in the cellar of the building caused the ISDN interface to crash at irregular intervals – re-establishing the line required asking the German Telekom to perform a manual reset.

The system prompts were recorded in a studio (male voice, high German dialect) and downsampled from 16 KHz/16 bit linear coding to 8 KHz/8 bit a-law coding, the European ISDN standard. The software to run an interview was built by incrementally extending the sample software that came with the audio board. Although the software performed well during the recordings, it has serious drawbacks: it is inflexible, i.e. any change in the specification of the interviews requires a recompilation, and it wastes hardware resources by not allowing more than 5 parallel recordings.

SPEAKER RECRUITMENT

There were two phases of speaker recruitment. In the first phase, a hierarchical recruitment scheme was applied. In the second phase, prompt sheets were distributed in university lectures and each caller was asked to pass on the prompt sheet to friends and relatives.

Hierarchical Recruitment

In the hierarchical recruitment scheme, department managers of Siemens AG were requested to ask 10 employees to call the speech server. Both the callers and the manager were given a telephone card (value 6.- DM).

In the hierarchical recruitment, approx. 6000 prompt sheets were distributed to approx. 300 persons at Siemens AG locations in Germany during August until October 1995. By the end of October, roughly 600 calls had been recorded (figure 2). Speakers recruited within Siemens continued to call during November and December although no more prompt sheets were distributed after October.

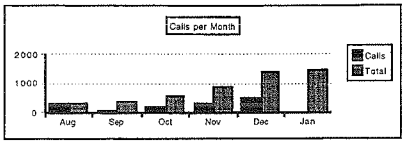

Figure 2. Calls per month

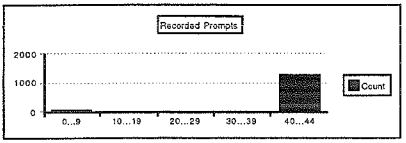

Figure 3. Item count

University recruitment

By mid-November, 60% of the callers were male, 40% were female. To balance the gender distribution, more female speakers were needed. In November, 1500 prompt sheets were distributed in large university lectures in the language sciences (which are attended mostly by female students). A telephone card (value 6.- DM) was promised to the first 50 female speakers who would call and return a complete personal information sheet.

Data collection and speaker statistics

The rate of response for the hierarchical recruitment within Siemens AG was 21.5%, for the university recruitment it was 13.2%. 89% of the calls were complete, i.e. had 40 and more recordings, 7% of the calls were aborted early due to technical or other problems (figure 3).

By mid-January 1996, 1488 calls had been recorded and a 50:50 male to female ratio for 1000 usable calls was achieved. Speakers were asked for two different geographic items: the place from which they called, and the federal state of Germany where they first entered school (the underlying assumption being that this age is decisive for the speaker's dialect). This information was used to determine the regional distribution.

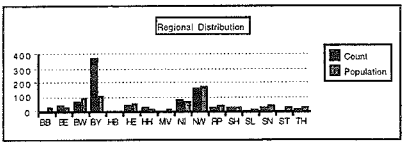

Figure 4. Regional distribution

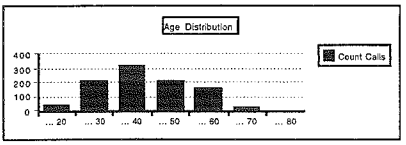

Figure 5. Age distribution

The regional distribution roughly matches the population distribution for the West German federal states with the exception of Bavaria (BY), where the number of speakers was more than double as high as necessary. For the East German states (BB, MV, SN, ST, TH) only half as many speakers as necessary have been recorded (figure 4). The over-representation of Bavaria is explained through the fact that Siemens is a Bavarian company and the university recruitment in Munich; the under-representation for the East German states is due to the fact that Siemens has only a few sites in these states.

The age distribution fulfills the SpeechDat(M) requirements of having at least 20% of the speakers in each of the age groups 17-30, 31-45, and 46-60 years of age (figure 5). If speaker gender is considered, then there is a slight under-representation of male speakers in the age group 17-30 (19%), and an under-representation of female speakers in the age group 46-60 (17%).

ORTHOGRAPHIC TRANSLITERATION

All SpeechDat calls are transcribed orthographically. For an efficient transcription, a simple transcription system was devised which contains only a very small set of markers to denote articulatory (e.g. coughing, laughing, lip smacks) and non-articulatory noise (line noise, handset noise, background noise), signal truncation, or incomplete words.

A pronunciation lexicon was established from the transcriptions, with the pronunciation denoted in the SAM phonetic alphabet. This dictionary contains approx. 3600 word types and approx. 192.000 word tokens.

A system to facilitate the orthographic transcription was developed in HyperCard on the Macintosh. This system displays a complete prompt sheet at once in the upper part of the screen, and the current item text and a transcription field in the lower part of the screen (figure 6). The speech signal is presented auditively only. The transcription guidelines are available on-line.

Since each prompt sheet could be used several times, the contents of the last transcription of a particular prompt sheet were available for the next transcription of the same sheet. This transcription software proved to be quite efficient: it took experienced transcribers approx. 12 minutes to transcribe a complete interview of approx. 4 minutes of speech.

The result of a transcription is a set of SAM header files which contain for each recorded item administrative data, i.e. file name, recording date and time, session number, etc., and descriptive data, i.e. item mnemonic, item text, item transcription, etc., in ASCII format.

VALIDATION

By mid-January 96, all calls had been transcribed. 1000 calls (500 male and 500 female speakers) were selected for distribution. The call directories containing the signal files and the SAM header files were transferred to three CD-ROMs according to the SpeechDat file system specifications.

SPEX (Speech Expertise Center) in Nijmegen, the Netherlands, was responsible for the validation of the SpeechDat(M) corpora. The German corpus was the first corpus collected according to the SpeechDat specifications to be turned in for validation, and hence served as a test for the validation procedure.

A number of issues were raised by this validation:

- documentation errors: missing contents file, missing data on internal validation
- lexicon errors: missing and superfluous entries, wrong word count
- transcription errors: transcription of empty files, spelling errors, inappropriate use of markers

The documentation errors and the lexicon errors have been corrected wherever possible. There is an ongoing discussion on the transcription errors: obvious errors such as misspellings, transcriptions of empty signal files, and formal errors have been corrected. Other reported errors, e.g. the use of markers to denote noise, are difficult to correct because transcription is always subjective. In table 3 the error rate for the transcriptions is divided into true transcription errors (typographic and formal errors), and errors concerning the use of noise markers. Long items are the connected digits, natural numbers, money amounts, spelled words, application phrases, and the sentences; everything else was considered to be a short item (van den Heuvel, 1996). The relatively high error rate for German is due to the fact that personal pronouns in German usually begin with upper case letters. In the transcription, only lower case personal pronouns were used, and this was counted as an error by SPEX because it was not mentioned in the documentation.

Language	Error% long items		Error% short items	
	Transcription	Background	Transcription	Background
English	5	48	5	19
French	6	46	2	33
German	8	36	4	44
Italian	2	28	1	25
Portuguese	2	7	3	7
Spanish	3	15	1	10
Swiss French	2	3	1	1

Table 3. Error rates for the SpeechDat(M) corpora

SUMMARY OF EXPERIENCES AND OUTLOOK

Speaker recruitment is the single most critical issue: the response rate is low, delays between sending out a sheet and receiving a call may be long (up to 8 weeks in SpeechDat(M)!), and it may be difficult to achieve proper age, region, and gender distribution. In SpeechDat II a monitoring procedure will be installed: calls will be checked for speaker age, region and gender immediately, speakers will be urged to call early, and the incentive for the intermediate distributors of prompt sheets will be proportional to the numbers of speakers they recruit.

An unproven technology – in 1995 ISDN primary rate interfaces were not common in Germany – led to technical problems which entailed a significant delay. After the technical problems had been solved, the setup worked reliably.

File administration was too complex: files were processed on three different platforms (SCO Unix for the recording, Macintosh for the transcription, and DOS for the production of CD-ROMs), and they were named differently in the recording directories and the final delivery database. This led to considerable file transfer delays (1 GB over 10Mb/s Ethernet takes $>10^4$ seconds if there is no other traffic on the net) and to inconsistencies between the recorded calls and their transcriptions, and the final CD-ROM distribution. In SpeechDat II, files will have the final distribution names and be stored in the proper file system hierarchy right away. Transcription and CD-ROM production will be on the Macintosh, with the recording platform file system mounted.

The use of an RDBMS for the generation of prompt sheets proved to be a flexible and efficient solution. The RDBMS was subsequently also used to store the transcription lexicon and the transcriptions themselves; the database structure is also suitable for SpeechDat II.

Hardware failures and human negligence do occur: there was ahead crash on the 5 GB external hard disk, and due to a programming error an entire hard disk partition containing 300 calls was deleted. Luckily, a reliable backup procedure had been installed.

These issues are currently being addressed for the SpeechDat II recordings which are scheduled to begin in October 96.

Information on the SpeechDat project, including the German SpeechDat II corpus specification (Draxler, 1996) and all public deliverables, can be found at

http://www.phonetik.uni-muenchen.de/SpeechDat.html

NOTES

(1) SpeechDat(M) was funded by the Commission of the European Union under contract LRE63314, SpeechDat II is funded by the Commission of the European Union under contract LE2-4001.

REFERENCES

den Os, E.A., Boogaart, T.I., Boves, L., Klabbers, E. (1995) *The Dutch Polyphone Corpus*, Proc. of Eurospeech '95, Madrid, 825-828

Draxler, C. (1996) *SpeechDat II Language Dependent Database Specifications for German*, SpeechDat LE2-4001 internal report, draft version

van den Heuvel, H., Sanders, E. (1996) *The SpeechDat(M) Validation: Results and Recommendations*, SpeechDat LE2-4001 - SD1.3.0 internal report

Winski, R. (1996) *Definition of Corpus, Scripts and Standards for Fixed Networks*, SpeechDat LE2-4001 - SD1.1.1 internal report