SpeechDat Experiences in Creating Large Multilingual Speech Databases for Teleservices

Christoph Draxler (1), **Henk van den Heuvel** (2) **& Herbert S. Tropf** (3)
(1) University of Munich [draxler@phonetik.uni-muenchen.de]
(2) SPEX, University Nijmegen [H.v.d.Heuvel@let.kun.nl]
(3) Siemens AG, Corporate Technology Department [Herbert.Tropf@mchp.siemens.de]

Abstract

In this article experiences in creating large multilingual speech databases for teleservices within a large consortium are reported in order to inspire, to facilitate or to compare the set-up and progress of other enterprises for collecting large speech databases. The focus will be on following aspects: Objectives, benefits, and strategy; project organization; database contents and creation; validation of the databases.

Objectives, benefits, and strategy

The main objectives of the SpeechDat project are

- the creation of large-scale speech databases for voice driven teleservices, and
- the coverage of all 11 official languages of the European Union and some major dialectal variants and minority languages.

In total, 28 databases are being collected: 20 databases are recorded over the fixed telephone network (FDB), 5 databases over the mobile network (MDB), and 3 databases are designed for speaker verification via telephone (SDB). The size of the databases ranges between 500 and 5000 calls by different speakers.
For the academic partners these spoken language resources are suitable for the improvement of speech processing technology, whereas industrial partners will use these resources today and in the near future to develop a number of applications including

- information services (e.g. timetable information);
- transaction services (e.g. home shopping, home banking);
- other call processing services (e.g. voice mail handling, call centre systems).

Three basic features of SpeechDat proved to be of great strategic value:

- extensive common specification of the content of the databases in order to cover a wide area of applications with practical relevance;
- coherent and consistent design of the database format in order to minimize cost for the development of multilingual teleservices;
- integration of a strict and thorough validation procedure carried out by a neutral institution in order to guarantee spoken language resources with sufficient quality.

Project organization

Consortium

The SpeechDat consortium consists of 12 contractors, 8 associated contractors, and several subcontractors. In principle, contractors are industrial companies and telecoms, whereas associated contractors and subcontractors are universities and public research institutes. (A detailed list of partners and corresponding databases being created is given in Höge *et al.*, 1997.)
For such a large consortium a considerable overhead is unavoidable for building up and maintaining the project infrastructure and flow of communication. In comparison to a small-scale project this disadvantage is easily compensated by the large number of resulting databases to which every partner will have access at the end of the project, since the conformity of all databases with respect to content, format, and quality will save a significant amount of cost to develop voice driven teleservices for several languages.

Structuring of work

The work within the project is being performed in 3 main phases. During the first phase the content, the standards and the dissemination procedures were defined which must be fulfilled by each producer of a SpeechDat database.
The second phase comprises the installation of a recording device, the prevalidation of small databases of 10 speakers in order to exclude severe errors during actual recordings, recruitment of speakers, creating appropriate speech files of the recordings, creating the corresponding annotation files, and documentation of the content of the databases. For databases containing more than 2000 calls a subset of the first 1000 calls has formally to be delivered. This separate step has two advantages: First, an intermediate validation can be performed in order to detect deviations which then can be corrected before all recordings are completed, and second, these database subsets can be made available to other SpeechDat partners and third parties in an early stage before the full databases are finished. These preliminary databases can e.g. be used to bootstrap speech recognizers for teleservices which are easy to handle.
The third phase covers the validation of those preliminary database containing 1000 calls, the validation of the full databases, and possibly one or more revalidations in case a database did not correspond to the specifications according to the main validation.

These milestones of prevalidation of small databases (10 calls), intermediate validation (1000 calls), main validation (full databases), and revalidation turned out to be a valuable means to structure the work load and to control the work progress. In fact, within SpeechDat the two sub-phases of building the recording platform including prevalidation, and the revalidation were somewhat underestimated with respect to their importance and work load involved. These sub-phases could have been established more explicitly and thus would have been controllable more efficiently.

Management and responsibilities

With respect to standards for projects funded by the European Union the management structure of SpeechDat is basically straightforward. Next to the main contract between the consortium and the Commission of the European Communities there exists a consortium agreement that defines certain rights and obligations of the parties in respect of the carrying out of the contract.
There is one project coordinator (Siemens AG), four work package managers (Matra Communications, Siemens AG, SPEX, Vocalis Ltd.), several task managers, and the representatives of the contractors. Since it is a rather large consortium a steering committee consisting of the coordinator and the work package managers is formal part of the management structure. Certain tasks are delegated by the representatives of the contractors to this steering committee. Its major task is to decide on the consequences of the validation reports, namely whether a database is officially accepted, to be corrected by its producer and revalidated, or definitely rejected. This procedure is also to be seen as a valuable and indispensable feature of SpeechDat in order to warrant the quality of its resulting databases.
By definition, the contribution of each (associated) contractor has the exchange value of a database of 5000 speakers recorded over the fixed telephone network (FDB5000). Thus, besides technical reports in principle each partner is obliged to deliver at least one FDB5000, or a set of smaller FDB and/or smaller MDB or SDB. The sort of database(s), the language(s) to be covered, the size of database(s), and also the responsibility for other deliverables was clearly determined for each single partner right from the commencement of the project, which helped to avoid a main source of potential differences among the consortium partners during the lifetime of the project.

Access and dissemination of results

Following basic agreements proved to be a reasonable and effective practice in order to handle the substantial rights and obligations of the partners, i.e. the access to and the delivery of databases. Particularly the first one can be motivating for speeding up the work progress:

- All databases generated in the project will be promptly available for use by each contractor provided his own databases are validated and accepted.
- Each (associated) contractor is allowed to supply to third parties only his own database(s).
- All databases generated in the project will be available for research purposes by each associated contractor at the end of the project.
- All databases are made available to third parties no later than 18 months after the end of project.
- Each (associated) contractor is obliged to offer his own databases(s) to ELRA for distribution.

Progress and efficiency

The planned lifetime of the project was 24 months starting in March 1996. It was clear from the beginning that this would be a very tight schedule for such a large consortium. Nevertheless, this duration was formally set to prevent the illusion of having "a lot of time" for carrying out the work. Even if this factor is taken into account time delay is the major concern of SpeechDat. The definition phase took 10 months where 6 months were planned, and the recording phase will take about 20 months instead of 13, mainly due to difficulties in recruiting sufficient speakers. In contrast to these delays the actual time for validation approximately corresponds to the original planning, namely four databases per month.
Unfortunately, only moderate and indirect means are available in SpeechDat to stimulate a more rapid work progress. In principal, the cooperation and contribution of partners are based only on good will of the single partners. Means of control that are actually used include rigid monitoring of planned and real figures of recordings and annotations per database on a monthly basis, every three months a workshop including detailed individual progress reports, access to other databases only as soon as the own databases are validated and accepted, and as *ultima ratio* retain or cut of funding.
Besides the delay in time all other main aspects of the project, like results of validation, expectations concerning the finalizing of all databases, or stability of the consortium can be judged quite positively.
End of March 1998 all 11 partial databases (FDB1000) were (re-)validated and accepted, 3 of 20 full FDB were (re-)validated and accepted, and 1 of 3 SDB was submitted for validation.
Since begin of the actual recording phase (January 1997) the monthly global recording rate was 5.8% on average of all 63,000 calls to be collected within the project. The corresponding monthly global annotation rate (begin March 1997) was 5.5% on average. On the date of reporting (March 1998) both rates have been relatively stable since 4 months at the level of about 7%, which means project-wide a total of about 4,500 calls and annotations per month, or about 160 calls and annotations per database per month.

Database contents and creation

Technical set-up

All SpeechDat databases are recorded on telephone servers connected to digital ISDN lines, either via base rate ISDN (BRI; 30 data and 2 command channels) or primary rate ISDN (PRI; 2 data and 1 command channel) interfaces. The signal format is 8bit 8KHz alaw, the European ISDN standard.
In all recording sites, speech servers handle incoming calls and guide the caller through an interview. The speaker is prompted for input by a beep, and recording duration is either fixed or determined by silence

detection. The speech servers store the recorded signal directly on the hard disk, one file per recording. Recording platforms are standard PCs or high-end workstations running either UNIX or Windows.
None of the recording platforms could handle all PRI data channels in parallel. During the course of SpeechDat low-cost BRI cards with a standardized communication application programming interface (CAPI) instead of proprietary libraries became available. This allowed scaling up to the required recording capacity simply by installing additional speech servers and leasing additional BRI lines, and it has made the exchange of speech server software within the project feasible.
One requirement of SpeechDat is that all calls be unique. In the FDB, multiple calls of one speaker calling from the same environment or using the same prompt sheet are considered as duplicate calls. In the MDB, a caller may call up to four times to cover all environments, so all calls in excess of these four are considered duplicate calls. At some recording sites, recording sessions were given unique session numbers. This allows a precise tracking of prompt sheets and also the continuation of recordings after technical failures but causes considerable administrative overhead and may lead to delays if speakers do not respond quickly. At other sites, consistency checks on the speaker and call data are performed to determine duplicate calls. These *a posteriori* checks can be performed during call monitoring or annotation, and thus they generally require some oversampling of recordings.

Data format

The major aim of SpeechDat is the exchange of data. This requires a precise specification of the signal and annotation data formats and the overall file system structure of the database.
It was decided to use the ISO 8859 family of alphabets for the annotation. ISO 8859 provides different code tables for all Western and Eastern European languages, and for Arabic and Hebrew.
For the signal and data format there were two options: the SAM file format (Tomlinson *et al.*, 1988), where signal and annotation data are held in at least two distinct files, and the NIST file format, where the annotation is included in a header in the signal file. From a data processing point of view, the SAM format is simple and elegant because it allows the modification of annotation files – for example updating and extending – without touching the signal file. Furthermore, because the signal files do not contain annotation headers, they can be output with any alaw-compatible audio output software, e.g. as a helper application in WWW browsers (Draxler, 1997).

```
LHD: SAM, 5.10
DBN: SpeechDat_German_Fixed_Network
VOL: FIXED1DE
SES: 0003
DIR: \FIXED1DE\BLOCK00\SES0003
SRC: A10003A1.DEA
CCD: A1
SHT: 4294-8
CMT: *** signal data ***
BEG: 0
END: 24000
REP: Dept. of Phonetics, University of Munich,
Germany
RED: 14/Dec/1997
RET: 18:14:00
SAM: 8000
SNB: 1
SSB: 8
QNT: A-LAW
CMT: *** speaker data ***
SCD: UNKNOWN
SEX: F
AGE: 30
ACC: BY
CMT: *** environment data ***
REG: UNKNOWN
ENV: OFFICE
NET: PSTN
PHM: TOUCH-TONE
LBD:
CMT: *** transcription data ***
LBR: 0,24000,,,,Nachricht
LBO: 0,12000,24000,[spk] Nachricht
ELF:
```

Figure 1: Sample SAM label file

The lexicon is rather simple. Each entry consists of the orthographic form of the item, an optional frequency count, and at least one canonic pronunciation in a SAM phonemic alphabet. Optional pronunciation variants may be included.

```
Abend 34      a: b @ n t
```

Figure 2: Sample lexicon entry

The file system hierarchy can be specified either content dependent or independent. In a content dependent structure, files are organized by some criterion in the data itself, e.g. speaker gender or region. This structure is often intuitive, but also inflexible. In a content independent structure files are organized by formal criteria, e.g. session number. In SpeechDat, a content independent file structure was chosen. This allowed a mechanical generation of file names according to the ISO 9660 file name restrictions for CD-ROMs, and it permitted storing all data in the final file structure during recording already. The file system hierarchy of SpeechDat is suitable for multiple databases on one physical medium (e.g. digital versatile disks DVD), and it can be extended to accommodate future databases.
Each SpeechDat CD-ROM contains the full set of annotation and documentation data, and a fraction of the speech signal data. To eliminate the inherent redundancy in the SpeechDat label files and to speed up access to the annotation data a relational DBMS can be used to store the SAM files.

Tools and procedures

A 5000 speaker SpeechDat database of 50 utterances each contains 250.000 signal files and the same number of SAM label files. To cope with this amount of data automatization is mandatory. The content independent file system hierarchy allows storing recording session directories in the final CD-ROM file system structure, thus avoiding the transfer of files.
Software tools were developed at various sites to speed up SpeechDat annotation. These tools all implement some consistency checking for the annotation, e.g. correct use

of marker symbols, spell checking, etc., and directly support editing SpeechDat annotations, e.g. through buttons or keyboard shortcuts for converting numbers, dates, etc. to the corresponding orthographic form.
At most sites, incoming calls were registered before being annotated. This separation of registration and annotation allowed monitoring the recording progress and computing signal properties for the annotation, e.g. begin and end of speech in the signal file, assessment of signal quality, etc. At one site the extensive use of such preparatory computations permitted the annotation of up to 6 calls per hour, whereas at the other sites 2 to 4 calls per hour could be annotated.

Database content definition

Three different types of SpeechDat databases were specified for the fixed telephone network (FDB), the mobile telephone network (MDB), and speaker verification respectively (SDB). These databases share a core of roughly 40 items (Winski, 1997; Kordi 1996; van Velden *et al.*, 1996):

II		Utterance description
500+	4000+	
2	2	isolated digit items
4	4	digit/number strings
1+	1+	natural number
1	1	money amounts
2	2	yes/no questions
3+	3+	dates
2	2	times
6	3	application keywords/key-phrases
1	1	word spotting phrase using embedded application words
5	5	directory assistance names
3	3	spellings
4+	4+	phonetically rich words
9	9	phonetically rich sentences
43+	40+	TOTAL

Table 1: SpeechDat FDB corpus contents definition

The number of items in each category is determined by the minimum number of items needed for the training of speech recognizers. For the phonetically rich sentences and words a maximum number of item repetitions was specified to obtain a high number of phoneme contexts for a good coverage of di- and triphones.
For the item categories with a rather fixed vocabulary, e.g. digits, money amounts, date expressions, randomized procedures were used to generate prompt sheet items.
The set of application words and expressions was defined in a three stage process: First, all partners were asked to propose functionalities that they considered important. This complete list was then circulated by e-mail among partners. Each partner voted for the twenty-five most relevant items. Finally, the twenty-five items with the highest score were selected for the common core, and each partner was asked to provide expressions for these functionalities in his own language.
Besides the common core vocabulary, each partner was free to record additional material for his own purposes. The full specification of the SpeechDat contents is publicly available at
`http://speechdat.phonetik.uni-muenchen.de/`

Demographic specifications To ensure a good coverage of the speaker population the following demographic criteria were specified (Senia, 1997):

1) 50% (± 2.5%) male and female speakers
2) all accent regions had to be covered proportionally
3) a minimum of 20% of the speakers from the age groups 16–30 and 31–45 years, and 15% in the age group 46–60 years
4) 2% of the FDB calls had to be from a public phone
5) 25% of the MDB calls had to be from either home, public place, street or moving vehicle environment

To meet these specifications either an oversampling approach or a close monitoring of incoming calls and returned data sheets was employed within the project.

Speaker recruitment For all partners speaker recruitment was the most difficult task, and it took longer than expected for most partners. The following four basic types of recruitment were used:

1) recruitment by a market research company
2) direct mailing
3) snowball recruitment
4) public calls for participation, e.g. in newspapers or the Internet

As an incentive to participate, either a lottery was organized or donations were given to charity.
Recruitment via a market research company seems to be the most efficient, but also the most expensive means of getting speakers. Generally, market research companies do not guarantee a number of callers, but only the number of contacts. One major problem when using a market research company is that its demographic criteria may not match those of the project (e.g. accent classification).
Direct mailing can be employed successfully if prompt sheets can be distributed cheaply. This is often the case in medium-size companies, where internal mail can be used.
In a snowball recruitment, people are asked to supply the address of potential callers. As a motivation they are offered an additional reward, e.g. a number of lottery tickets proportional to the number of people recruited. This recruitment scheme causes some administrative overhead, but return rates are much higher than with direct mailing.
In public calls for participation, e.g. via newspapers or the Internet, people are asked to apply for prompt sheets. The rate of return is very low, but this is compensated for by the size of the audience reached. Depending on the cooperation of newspapers, calls for publication can be placed free of charge. Furthermore, regional newspapers allow a fine control of regional coverage. The Internet is another means of publishing calls for participation; here, online prompt sheets can be provided directly to callers.
It can be said that all recruitment schemes except marketing research have to be employed in parallel. Furthermore, a speaker database which contains speaker addresses as well as demographic data is of immense value for any speaker recruitment.

Validation of the databases

In this section we explain in some detail the validation of the SpeechDat(II) databases in terms of: objectives, evaluation protocol and experiences so far.

Validation objectives

The SpeechDat project is featured by a thorough validation protocol. The specifications which the databases should meet are evaluated by an independent validation centre, SPEX. SPEX is the primary validation centre for speech databases within ELRA. This approach warrants that each database that is produced by the consortium is in agreement with a well-defined set of minimum quality standards.

The validation procedure proceeds in four steps:

1) Prevalidation of a small database of 10 speakers. The objective of this stage is to detect serious errors before the actual recordings start.
2) Intermediate validation on the first 1000 calls of databases larger than 2000 calls.
3) Validation of complete databases. The database is checked against the SpeechDat specifications and a validation report is generated.
4) Revalidation of complete databases. In case the validation report shows that improvements of a database are necessary or desirable, then (part of) the database can be offered for a second validation, and the validation report is updated.

The final validation report is put onto the final CDs as part of the SpeechDat database.

Validation protocol

More in detail, SPEX checks the following database features:

- Database design, in terms of recorded items;
- Structure of database directories and file names;
- Completeness of the database in terms of missing or unusable files;
- Completeness of the documentation; acoustic quality of speech files in terms of clipping rate, SNR, and file duration;
- Contents and format of the annotation files;
- Transcription quality;
- Lexicon format and completeness;
- Speaker characteristics in terms of sex, age, and dialect region;
- Environmental conditions;
- Training and test protocol.

For each of these checks a set of validation criteria was formulated by the consortium (van den Heuvel, 1997).

A lot of work can be done automatically. Software was written to check file formats, internal consistency, missing files, transcription symbols used, speaker and environment balances, etc. The interpretation of the software output involves human interference, as does the editing of the validation report. Furthermore the evaluation of the transcriptions and of the database documentation is done by hand. Evaluation of a database now takes a minimum of two weeks. In these two weeks two validations can be run in parallel.

Lessons learnt for validation

When this paper was written, a validation on the first 1000 speakers of each of 11 larger databases (>2000 speakers) for the fixed network had been carried out, and some of our experiences are worth reporting.

First of all, problems were encountered in speaker recruitment, data formatting and CD burning on the side of the database producers which caused delays in the delivery of the database to the validation centre. For this reason the schedules for delivery and validation of databases in a project of this size regularly undergoes major changes. For the validation centre this involves a continuous rescheduling of database validation for SpeechDat and of other activities.

Upon arrival of a database, we learnt to pay attention to a set of notoriously problematic files first. These files are the file with main documentation (often incomplete), the lexicon (does often not contain all words in the transcriptions), and the file with acoustic measures of the speech files (often absent). In case of incompleteness of the lexicon it must be checked first if the missing entries are really missing in the lexicon or if they are misspellings in the transcriptions, in which case the lexicon is not to blame but the transcriptions.

The above mentioned files are the first to be checked so that in case of absence or deficiencies an e-mail can be sent out to the database producer. The producer can then work at updates of the files and send these to the validation centre, whilst the rest of the validation procedure is still in progress. In this way further delay is minimised.

The project takes the rule that each item of the database must be effectively complete up to 95%. This means that a maximum of 5% of the speech files of each database item (e.g., isolated digits, or application words) may be absent or only contain noise according to their transcriptions. It appears that this criterion is, for obvious reasons, most difficult to meet for short (single word) items and for items at the back of the prompt sheet, which run the risk of premature call-interrupts.

The need is felt for a better evaluation of acoustical quality than could be realised in the present project. In this project the acoustical measures are just calculated and presented to the client who can then select sessions according to his own training and test desires. But on reconsideration it seems better to perform an extra auditory check on calls with extreme values for e.g. SNR, with the possible rejection of such calls, or at least the requirement to put such calls in the test set partition of the database.

The evaluation of the orthographic transcriptions is not straightforward. A lot of factors are involved which effect the outcome of the evaluation. This is especially true for the transcription of the noise symbols. SpeechDat has 4 symbols for non-speech acoustic events: [fil] for filled pauses; [spk] for speaker noises; [sta] for stationary noises; and [int] for intermittent noises (Senia & van Velden, 1997). The specifications state (only) that these noises are not transcribed if they are low-level and non-intrusive. But still whether or not such noises are transcribed depends on the exact instructions and intentions of the database producers, the subjective impressions of the transcribers, and the working environment of the transcribers (use of head set, noise in the room). And the same sources of variance are encountered at the site of the validation centre. As a general rule, we use native

speakers for transcription validation, we follow the instructions of the producer and further accredit the given transcription the benefit of the doubt.

As could be expected, the transcription error percentages we observe on samples of about 2000 utterances per database vary a lot: between 1.2% and 5.2% for speech, and between 0.1% and 15.0% for non-speech (noise).

For SpeechDat the maximum allowed percentage of utterances with an error in the transcription is 5% for speech, and for non-speech the permitted maximum is 20%. Since the omission of a noise symbol is judged an error during validation whereas the insertion of a noise symbol is not, a low error rate for the transcription of non-speech does not imply that the noise symbols are exactly located where the validation centre would put them. It is just the reflection of a conservative stance that any utterance without such a symbol is most probably "clean" of noises.

Our experience with the prevalidation phase is very good. By requiring that a mini-database of 10 speakers is checked first, it was avoided that some partners forgot to record a mandatory item in the full database, used wrong transcription symbols, wrong SAMPA phoneme symbols in the lexicon, etc. Further, it enabled the validation centre to develop and test the validation software before the actual database validation had to start. Similarly, this allowed the producer to develop the software for database creation and formatting in an early phase of the project.

Outlook

The basic strategies in carrying out such a project for collecting large speech databases are currently adopted by the project SpeechDat(Car) for in-car voice based applications, and by the project SALA (SpeechDat across Latin America) for collecting Spanish and Portuguese telephone speech databases to cover all major language variants of Latin America, and will also be taken over in the near future by the project SpeechDat(E) for creating telephone speech databases of several Eastern European languages. A starting point for further information and cross-references is

`http://speechdat.phonetik.uni-muenchen.de/`

References

Draxler, Chr. (1997). WWWTranscribe – A Modular Transcription System Based on the WWW. In *Eurospeech '97*, Rhodos.

Heuvel, H. van den (1997). Validation Criteria. *SpeechDat Technical Report SD1.3.3.*

Höge, H., H.S. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach & K. Choukri (1997). European Speech Databases for Telephone Applications. In *ICASSP'97, Munich.*

Kordi, K. (1996). Definition of Corpus, Scripts, and Standards for Speaker Verification. *SpeechDat Technical Report SD1.1.3.*

Senia, F. (1997). Specification of Speech Database Interchange Format. *SpeechDat Technical Report SD1.3.1 (Version 4.4).*

Senia, F. & J.G. van Velden (1997): Specification of orthographic transcription and lexicon conventions. *SpeechDat Technical Report SD1.3.2.*

Tomlinson, M., R. Winski, W. Barry (1988). Label file format proposal. *Esprit project 1542 (SAM): Extension Phase, Final Report.*

Velden, J.G. van, D. Langmann, M. Pawlewski (1996). Specification of Speech Data Collection over Mobile Telephone Networks. *SpeechDat Technical Report SD1.1.2/1.2.2.*

Winski, R. (1997). Definition of corpus, scripts and standards for Fixed Networks. *SpeechDat Technical Report SD1.1.1.*

Acknowledgements

Part of the SpeechDat project is funded by the Commission of the European Communities, Telematics Applications Programme, Language Engineering, Contract LE2–4001.