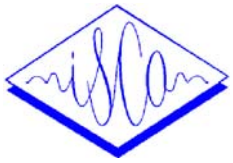

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEECHDAT MULTILINGUAL SPEECH DATABASES FOR TELESERVICES: ACROSS THE FINISH LINE

Harald Höge(1), Christoph Draxler(2), Henk van den Heuvel(3), Finn Tore Johansen (4), Eric Sanders (3), Herbert S. Tropf (1)

(1) Siemens AG, Corporate Technology Department, Munich, Germany;
(2) Ludwig-Maximilian University Munich, Germany;
(3) SPEX, Nijmegen, Netherlands;
(4) Telenor R&D, Kjeller, Norway

Harald.H.Hoege@mchp.siemens.de

ABSTRACT

The goal of the SpeechDat project is to develop spoken language resources for speech recognisers suited to realise voice driven teleservices. SpeechDat created speech databases for all official languages of the European Union and some major dialectal varieties and minority languages. The size of the databases ranges between 500 and 5000 speakers. In total 20 databases are recorded over the fixed telephone network, 5 databases over the cellular network, and 3 databases are designed for speaker verification. To date the project has successfully reached its end. This paper briefly describes the project, addresses the validation of the databases, their availability to consortium members and third parties, publicity and awareness, and the spin-off of the project in speech recognition research.

10.21437/Eurospeech.1999-495

1. INTRODUCTION

For current speech recognition technology, the availability of spoken language resources (SLR), i.e. speech databases, pronunciation lexica and text corpora, is crucial [17]. These SLR are language specific and have to be tuned to the specific application area. The goal of the SpeechDat project [18,19] is to create speech databases to train speaker-independent recognisers for all official languages in the European Union. Such recognisers can be used for any voice driven teleservices and can be accessed via the fixed and the cellular network. SpeechDat is the first project in which commercially usable speech databases have been produced within an international consortium of industrial and academic partners. This consortial approach was chosen because it allows the free exchange of equivalent databases which are otherwise expensive and time-consuming to produce for each partner individually.

A crucial issue was the specification of the databases. The databases should be designed to train several special-purpose recognisers (e.g. recognition of isolated command words, digit strings, numbers, dates, continuous speech). Further, the databases should cover the various speaker-related, environmental, and transmission characteristics. Additionally, the project was restricted in terms of cost and time. Given these demands and constraints the optimal design of the databases had to be primarily gained by the project itself: There were no comparable databases where these issues could be studied. Consequently the specification of the SpeechDat databases was based on the best knowledge and 'feeling'. As the databases are ready now and the first teleservices have been set into the field all these open questions can now be investigated. Due to the commercial success of the distribution of the SpeechDat databases via the European Language Resource Distribution Agency ELDA [28] and due to the many successor projects building the 'SpeechDat Family' [29] the chosen specification seems to be a solid basis for training all the different types of recognisers.

2. BASICS OF THE PROJECT

The main result of the project are 28 databases each containing between 500 and 5000 annotated calls (cf. Table 1):

- 20 databases recorded over the fixed telephone network (FDB)
- 5 databases recorded over the mobile network (MDB)
- 3 databases designed for speaker verification (SDB)

They cover all official languages of the European Union and some major dialectal varieties and minority languages. Each database comes with an orthographic transcription for each speech file and a lexicon which contains a canonical phoneme transcription of each word in the transcriptions.

With some tolerance the following demographic criteria are met with respect to the selection of speakers:

- gender: 50% male and female
- age: min. 20% 16–30 years, min. 20% 31–45 years, min. 15% 46–60 years
- region: all accent regions covered proportionally

Recordings were made from different environments: For FDB two environments were distinguished: home-office, and public place. For MDB and SDB, four environments were defined: home-office, public place, along a busy street, and moving vehicle.

All of the databases have a common core of recorded utterances (cf. Table 2) and a consistent design of the

format which facilitates the development of teleservices in several languages considerably [21,22,23,24].

Table 1: Overview of all SpeechDat databases

DB-ID	Type	Language (variant)	# calls	# calls per speaker
1	FDB	Danish	4000	1
2	FDB	Flemish	1000	1
3	FDB	French (Belgium)	1000	1
4	FDB	German (Luxembourg)	500	1
5	FDB	French (Luxembourg)	500	1
6	FDB	English (UK)	4000	1
7	SDB	English (UK)	2400	20
8	FDB	Welsh	2000	1
9	MDB	English (UK)	1000	1
10	FDB	Finnish	4000	1
11	FDB	Swedish (Finland)	1000	1
12	FDB	French	5000	1
13	SDB	French	2400	20
14	MDB	Dutch	1000	4
15	FDB	French (Switzerland)	3000	1
16	FDB	German (Switzerland)	2000	1
17	SDB	French (Switzerland)	1000	50
18	FDB	German	4000	1
19	FDB	Slovenian	1000	1
20	FDB	Greek	5000	1
21	FDB	Italian	3000	1
22	MDB	Italian	1000	4
23	FDB	Portuguese	4000	1
24	FDB	Spanish	4000	1
25	FDB	Swedish	5000	1
26	MDB	Swedish	1000	1
27	FDB	Norwegian	1000	1
28	MDB	German	1000	1

The SpeechDat databases are recorded on telephone servers connected to ISDN lines. The signal format is 8 bit, 8 kHz, A-law.

For each of the 11 FDBs with more than 2000 speakers an additional database was created which consists of a subset of 1000 calls. These were used within the consortium as exchange material for partners who made databases of a similar size (e.g. the Norwegian FDB). Some of these 1000 speaker databases are made public via ELRA [28].

A project partner can use all SpeechDat databases which are not produced by him for exploitation, but he cannot use it for commercialisation. From 1 July 2000 at the latest all SpeechDat databases have to be made publicly available for exploitation (though not for commercialisation). At present, all databases are finished and some of them are already being distributed by ELRA.

In order to give a rough idea of the production cost: The planned total cost of the project was 3.3 MECU but this did not sufficiently cover the actual cost.

Table 2: Database contents

Utterance description	# per call
Isolated digit items	2
Digit/number strings	4
Natural number	1+
Money amounts	1
Answers to yes/no questions	2
Dates	3+
Times	2
Application keywords/key-phrases	3+
Word spotting phrase using embedded application words	1
Directory assistance names	5
Spellings	3
Phonetically rich words	4+
Phonetically rich sentences	9
TOTAL	40+

3. PRODUCTION

In SpeechDat, the single most critical issue turned out to be speaker recruitment, and this was the reason for most of the delays experienced in SpeechDat. The following recruitment strategies were used [27]:

1. A market research company was charged with recruiting speakers. This approach is the most expensive, but it guarantees within a given time span a speaker population that complies with the requirements.
2. Speaker recruitment within a company was highly successful for some partners, and less successful for others. The Norwegian and Portuguese speakers were recruited mainly within the SpeechDat partner company; here, the companies proved to be sufficiently large to meet the demographic criteria of the speaker population. For other databases, e.g. the fixed network German DB, the internal recruitment was less successful; here the rate of response was less than 10%.
3. Calls for participation were published in newspapers, magazines, or on the Internet. People interested in participating were sent the prompt sheets. The rate of response varied considerably. Only very few callers could be recruited via the Internet. Paid magazine advertisements are very expensive; however, for well-targeted audiences, e.g. clients of a mobile network provider, such advertisements were a good way to start a database collection. Daily newspapers were often interested in publishing articles about the project. Such an article contains a phone number to apply for prompting material. Using newspapers with a regional distribution allowed the targeted collection of speech from specific regions.
4. In a snowball system, speakers are asked to recruit further speakers. The recruiter would receive an extra incentive, usually proportional to the number of speakers recruited, e.g. additional lottery tickets.

In all recruitment schemes speakers were offered an incentive to participate, e.g. a telephone card or a lottery ticket.

In SpeechDat annotation was purely orthographical with mispronunciation, noise and signal truncation markers. Annotations were performed by trained transcribers, usually phonetics or language science students. The annotation of an entire call of approx. 3.5 minutes speech took about 20 to 30 minutes. The annotation of spontaneous items naturally is slower, especially for utterances longer than 5 seconds. To speed up the annotation, some tools present the original prompt text to the transcriber so that this text had to be edited only; some tools feature editing buttons that perform often needed conversion tasks, e.g. conversion of digits to strings. Also, the use of off-line signal processing, e.g. to determine begin and end of speech, made the annotation more efficient. Finally, a consistency checker for the annotations allowed only formally correct annotations to enter the label files.

4. VALIDATION

SpeechDat followed a unique evaluation campaign in order to assure that all databases meet the specifications that were originally set up. Unique in the sense that an independent organisation checked all databases within and thus as part of the project itself.

The following aspects of a database were checked and compared to the validation criteria as agreed by the consortium: completeness and correctness of documentation; compliance to the database format specifications; completeness of recordings; correctness of the distributions of individual items; quality of the speech signals; balances of speaker and environmental distributions; completeness of the lexicon; quality of the orthographic transcriptions (checked by a native speaker of the language). The exact validation criteria for the databases, grouped for database class (FDB, MDB, SDB), are listed in [25].

The approval of a database for the SpeechDat consortium was not determined by the validation centre but by the Steering Committee of the project on the basis of the validation report edited by the validation centre. A database was re-validated if the consortium or the producer considered it necessary that (part of) a database be rectified. Table 3 shows the number of databases that were accepted in the first pass. All databases which were not approved were corrected and offered for revalidation. As a consequence, all originally envisaged databases are produced and meet the SpeechDat quality standards.

Table 3: Number of databases accepted after validation

	FDB	MDB	SDB
Accepted	16	2	0
Revalidation needed	2	2	1
Under validation	2	1	2
TOTAL	20	5	3

In general, MDBs and SDBs had more difficulties to pass the validation than FDBs, the reason being that more criteria had to be met, e.g. the number of calls per speaker and the stricter environmental conditions.

Not all original validation criteria could be maintained. For some checks the high number of databases failing the test indicated that there was something wrong with the criterion rather than with the databases. Thus in the course of the project, three criteria were revised. These pertained to: 1. The minimum number of tokens per phone in the phonetically rich words; 2. the compensation of missing files by other items in the database; 3. The maximum number of missing files for SDB.

5. EVALUATION OF THE USE OF THE SPEECHDAT DATABASES

The main application of the SpeechDat databases is the development of telephone speech recognition and verification systems. Such development is indeed taking place, both among commercial recogniser manufacturers and in research laboratories. A number of research results have already been published, e.g. in language identification [8], multilingual recognition [2,3], speaker verification [16] and general acoustic-phonetic modelling and adaptation for different environments and tasks [4,5,9,10,11,12,14,15]. Apart from this, the SpeechDat databases also represent a valuable collection of dialects and speakers for corpus-based linguistic and phonetic studies [1,6,7,13].

So far, most of the work published has been based on a precursor project called SpeechDat(M) [29]. The SpeechDat design differs from this, by the number of speakers and the languages covered, by the addition of mobile network and speech verification material, and by the improved phonetic coverage, especially in the isolated word corpora. In [12], it is shown that a straightforward HTK-based phonetic recogniser trained on a SpeechDat FDB1000 achieves reasonably good results (e.g. 14.3% errors on a 1100 “city name” recognition task), and that the phonetically rich isolated word and name material contributes significantly to the recognition performance.

Within the COST Action 249 "Continuous speech recognition over the telephone", a cooperative effort is being made to create a common, flexible vocabulary recogniser design based on SpeechDat databases and the HTK toolkit, using a fully automatic and language-independent training procedure. Preliminary test results are available for a few languages (Norwegian, Swiss German, Slovenian, English and Swedish), and show that a language-independent design is indeed feasible using the information present in SpeechDat. From the results obtained so far, error rates for an isolated digit task are 2.6% for Swedish, 2.3% for Norwegian and 4.2% for Slovenian. On a 30 “application word” recognition task, 1.5% errors have been obtained for Swedish, 4.9% for Norwegian and 0.9% for Swiss

German. More results on SpeechDat recognition are expected as the databases now become available to the research community.

6. CONCLUSION

The SpeechDat project can be considered a success. All 18 original partners crossed the finish line in an atmosphere of good cooperation and with the determination to get the best out of it. As a result 28 high quality speech corpora for 21 different language varieties have been created. The SpeechDat formula is prolonged in a number of successor projects: SpeechDat Car (aiming at wideband and GSM recordings in the car [26]), SpeechDat(E) (FDBs for five central and eastern European languages), and SALA (SpeechDat Across Latin America) [20].

It was further decided to register SpeechDat as an Internet domain and to continue to maintain the SpeechDat WWW server at http://www.speechdat.org. All publicly available specifications and reports can be found on this server, and all SpeechDat-related projects can be accessed from there.

Within the SpeechDat consortium a procedure for error correction was defined. Users of the databases are encouraged to report noted errors to our Web site. At (irregular) intervals update patches will be created and released based on these error reports.

Furthermore, a demonstration CD-ROM was produced which contains all public reports and samples from all speech databases. Finally, it is planned to organise an international workshop on the experiences gained with SpeechDat and similar databases in the Spring of 2000.

ACKNOWLEDGEMENT

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

REFERENCES

[1] C. Draxler: A multi-level description of date expressions in German telephone speech, *Proc. ICSLP 96*, pp. 1906-1909.

[2] U. Bub, J. Köhler, B. Imperl: In-service adaptation of multilingual hidden-Markov-models, *Proc. ICASSP 97*, pp. 1451-1454.

[3] J. Köhler: Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks, *Proc. ICASSP 98*, pp. 417-420.

[4] J. Junkawitsch, H. Höge: Keyword verification considering the correlation of succeeding feature vectors, *Proc. ICASSP 98*, pp. 221-224.

[5] U. Bub, H. Höge: Boosting long-term adaptation of hidden-Markov-models: Incremental splitting of probability density functions, *Proc. ICASSP 98*, pp. I-429-432.

[6] K. Kvale, A.F. Foldvik: "Four-and-twenty, twenty-four". What's in a number?, *Proc. Eurospeech 97*, pp 729-732.

[7] C. Draxler, S. Burger: Identification of regional variants of high German from digit sequences in German telephone speech, *Proc. Eurospeech 97*, pp. 747-750.

[8] D. Caseiro, I. Trancoso: Spoken language identification using the SpeechDat corpus, *Proc. ICSLP 98*, pp. 3197-3200.

[9] A. Nogueiras-Rodriguez, J.B. Mariño: Task adaptation of sub-lexical unit models using the minimum confusability criterion on task independent databases, *Proc. ICSLP 98*, pp. 2983-2986.

[10] L. Bahl et al: A method for modeling liaison in a speech recognition system for French, *Proc. ICSLP 98*, pp. 2447-2450.

[11] A. Kellner, B. Rueber, H. Schramm: Using combined decisions and confidence measures for name recognition in automatic directory assistance systems, *Proc. ICSLP 98*, pp. 2859-2862.

[12] F.T. Johansen: Phoneme recognition for the Norwegian SpeechDat(II) database, *Proc. ICSLP 98*, pp. 333-336.

[13] A.K. Foldvik, K. Kvale: Dialect maps and dialect research; Useful tools for automatic speech recognition?, *Proc. ICSLP 98*, pp. 153-156.

[14] A. Fischer, V. Stahl: Database and online adaptation for improved speech recognition in car environments. *Proc. ICASSP 99*, pp. 445-448.

[15] J.B. Mariño, P. Pachès-Leal, A. Nogueiras: The demiphone versus the triphone in a decision-tree state-tying framework, *Proc. ICSLP 98*, pp. 2463-2466.

[16] H. Melin, J.W. Koolwaaij, J. Lindberg, F. Bimbot (1998): A comparative evaluation of variance flooring techniques in HMM. *Proc. ICSLP 98*, pp. 1903-1906.

[17] Höge. H. (1998) Spoken Language Resources for Voice Driven Man Machine Interfaces. *Proc. LREC 98*, Granada, pp. 209-216.

[18] Höge, H., Tropf, H.S., Winski, R., Van den Heuvel, H., Haeb-Umbach, R. & Choukri, K. (1997) European speech databases for telephone applications. *Proc. ICASSP 97*, Munich, pp. 1771-1774.

[19] Draxler, C., Van den Heuvel, H., Tropf , H. (1998) SpeechDat Experiences in creating large multilingual speech databases for teleservices. Proc. *LREC 98*, Granada, pp 361-366.

[20] Moreno, A., Höge, H., Koehler, J., Marino, J. (1998) SpeechDat Across Latin America. Project SALA. *Proc. LREC 98*, Granada, pp. 367-370.

[21] Senia, F. (1997). Specification of speech database inter-change format. *SpeechDat Technical Report SD1.3.1* .

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

[23] Kordi, K. (1996). Definition of corpus, scripts, and standards for Speaker Verification. *SpeechDat Technical Report SD1.1.3.*

[24] 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.*

[25] Van den Heuvel, H. (1997). Validation criteria. *SpeechDat Technical Report SD1.3.3* .

[26] Van den Heuvel, H., Boudy, J. Comeyne, R., Euler, S. , Moreno, A. Richard, G. (1999) The SpeechDat-Car multiligual speech databases for in-car applications: some first validation results. *Proc. Eurospeech 99, Budapest.*

[27] Lindberg, B., Comeyne, R., Draxler, Ch., Senia, F. (1998) Speaker recruitment methods and speaker coverage. Experiences from a large multilingual speech database collection. *Proc. ICSLP 98*, pp. 2731-2734.

[28] ELRA/ELDA: http://www.icp.grenet.fr/ELRA/

[29] SpeechDat Family: http://www.speechdat.org/