EUROPEAN SPEECH DATABASES FOR TELEPHONE APPLICATIONS

Harald Höge (1), Herbert S. Tropf (1), Richard Winski (2),
Henk van den Heuvel (3), Reinhold Haeb-Umbach (4), Khalid Choukri (5)

(1) Siemens AG, Germany; (2) Vocalis Ltd., United Kingdom;
(3) SPEX, The Netherlands; (4) Philips GmbH, Germany; (5) ELRA, France
Harald.Hoege@mchp.siemens.de

ABSTRACT

The SpeechDat project aims to produce speech databases for all official languages of the European Union and some major dialectal variants and minority languages resulting in 28 speech databases. They will be recorded over fixed and mobile telephone networks. This will provide a realistic basis for training and assessment of both isolated and continuous-speech utterances, employing whole-word or subword approaches, and thus can be used for developing voice driven teleservices including speaker verification. The specification of the databases has been developed jointly, and is essentially the same for each language to facilitate dissemination and use. There will be a controlled variation among the speakers concerning sex, age, dialect, environment of call etc. The validation of all databases will be carried out centrally. The SpeechDat databases will be transferred to ELRA for distribution. Next databases to be recorded will cover East European languages.

1 INTRODUCTION

Speech databases are needed to train and test recognition and speaker verification systems. The SpeechDat project [1] funded by the Commission of the European Communities aims to produce speech databases for all official languages of the European Union and some major dialectal variants and minority languages. SpeechDat databases are focussed on the telephone environment including fixed and mobile telephone networks. The SpeechDat databases are designed in the spirit of the Polyphone databases proposed by COCOSDA and collected by the Linguistic Data Consortium. Yet the scope of the SpeechDat databases is much larger with respect to the number of databases and application areas, which include fixed and mobile telephone network and speaker verification over fixed and mobile telephone networks:

- Fixed Telephone Network: 20 databases, 500-5000 speakers each, 1 call per speaker
- Mobile Telephone Network: 5 databases, 250-1000 speakers each, 1-4 calls per speaker
- Speaker Verification - Fixed and Mobile Networks: 3 databases, 20-120 speakers each, 20-50 calls per speaker

and with respect to following extensions:

- Coverage of language variants and minority languages, e.g. Swiss and Belgian French, Finnish Swedish, Flemish, Welsh
- Validation and distribution mechanisms

The basis of the SpeechDat project was laid by the predecessor project SpeechDat(M), where 8 speech databases were produced. In SpeechDat(M) experiences were made including the following items:

- specification of content
- speaker recruitment, recording and transcription
- validation
- distribution
- scientific and commercial relevance

In the following these items will be sketched for SpeechDat after a short overview is given. Then the scientific and commercial relevance of the speech databases is sketched.

2 THE SPEECHDAT PROJECT

2.1 Overview

The 2 year project started in March 1996. The consortium consists of European partners mainly active in the telecom sector. Prime contractor is Siemens AG. The total cost of the project is about 3.3 MECU which is funded by about 2.0 MECU within the Language Engineering Cluster of the Telematics Applications Programme. Each partner is responsible for the production of its own databases. Following major milestones are set: (1) August 1996: Specification of content and standards of speech databases finished; (2) February 1997: Recording and validation of the first 1000 speakers of the databases finished; (3) February 1998: Validation of all databases finished. Validation is

done by the validation center SPEX [2]. The databases will be distributed via ELRA [3].

Differentiated in 3 types of databases the tables 1 to 3 below show in detail the languages covered, the partners who are responsible for producing the databases, the number of speakers to be recorded, and the number of calls per speaker if speakers have to be recorded more than once.

Table 1: SpeechDat databases concerning Fixed Telephone Network

Language (variant)	*Responsible Partner(s)*	*Speakers*
Belgian French	Lernhout & Hauspie Speech Products N.V.	1000
British English	GPT Ltd., GEC-Marconi Material Technology Ltd.	+[a] 4000
Danish	Aalborg University	+[a] 4000
Finnish	Tampere University. of Technology	4000
Finnish Swedish	Tampere University. of Technology	1000
Flemish	Lernout & Hauspie Speech Products N.V.	1000
French	MATRA Communication	+[a] 5000
German	Siemens AG, Universität München	+4000
Greek	Knowledge S.A., University of Patras	5000
Italian	Centro Studi e Laboratori Telecommunicazioni S.p.A.	+3000
Luxembourgish French	Lernhout & Hauspie Speech Products N.V.	500
Luxembourgish German	Lernhout & Hauspie Speech Products N.V.	500
Norwegian	Telenor AS	1000
Portuguese	Portugal Telecom S.A., INESC	+4000
Slovenian	University of Maribor	1000
Spanish	Universitat Politecnica de Catalunya	+[a] 4000
Swedish	Kungl Tekniska Hogskolan	5000
Swiss French	IDIAP, Swiss Telecom PTT	+[a] 3000
Swiss German	IDIAP, Swiss Telecom PTT	1000
Welsh	British Telecommunications plc	2000

a)Throughout this table "+" indicates that a corresponding database of 1000 speakers produced within the predecessor project SpeechDat(M) already exists. In the case of British English the owner of the SpeechDat(M) database is GEC Marconi Secure Systems Ltd. Tele Danmark is the owner of the Danish SpeechDat(M) database, Philips GmbH the owner of the French, Vocalis Ltd. the owner of the Spanish, and Swiss Telecom PTT the owner of the Swiss French SpeechDat(M) database.

2.2 Specification

The aim of SpeechDat is to produce databases to aid rapid development of voice driven teleservices in 20 European languages and major dialects or minority languages. It will provide a realistic basis for training and assessment of both isolated and continuous-speech utterances, employing whole-word or subword approaches. The specification of the databases for the fixed telephone network [4, 5], the mobile telephone network [6], and for speaker verification [7, 8] has been developed jointly, and is essentially the same for each language to facilitate dissemination and use [9, 10]. These specifications can be found in [1].

Table 2: SpeechDat databases concerning Mobile Telephone Network

Lang. (variant)	*Responsible Partner*	*speakers*	*calls/ speaker*
British English	British Telecommunications plc	1000	1
Dutch	Philips GmbH	250	4
German	Vocalis Ltd.	1000	1
Italian	Centro Studi e Laboratori Telecomm. S.p.A.	250	4
Swedish	Kungl Tekniska Hogskolan	1000	1

Table 3: SpeechDat databases concerning Speaker Verification

Lang. (variant)	*Responsible Partner*	*speakers*	*calls/ speaker*
British English	GPT Ltd., GEC-Marconi Material Technology Ltd.	120	20
French	MATRA Communication	120	20
Swiss French	IDIAP, Swiss Telecom PTT	20	50

The project provides a unique opportunity to record large corpora of coherently designed material in the major European languages. Each of the fixed network, mobile network and speaker verification databases are designed to address their respective goals, but also to complement each other. Thus the fixed and mobile network databases include large amounts of speaker verification impostor testing material, and the speaker verification database records speaker-dependent material useful to train and test repertory dialling. The use of common material in all three databases facilitates comparative testing of network factors.

The recorded material comprises immediately useable and relevant speech, including digit and letter sequences, numbers and money amounts, common application keywords and wordspotting phrases, dates, times, yes/no

responses, person and city directory assistance names, and phonetically rich words and sentences. Utterances are mainly read speech but include some spontaneous speech.

2.3 Speaker recruitment, recording and transcription

In SpeechDat the telephone callers are recruited via market companies or internally within institutions. Each speaker is provided with a prompt sheet. The prompts consist of questions to be answered (generating spontaneous speech) and text to be read (generating read speech). The set of prompt sheets are designed to provide a large variety of different tokens for each item category and in general many different prompt sheets have to be produced. Approximately 45 utterances are elicited per speaker which corresponds roughly to 10 minutes recording time. An instruction sheet accompanies the prompt sheet, and acoustic prompting and instructions guide the caller through the recording process.

Speakers to be recruited for a SpeechDat database have to fulfill several requirements (or in some cases mandatorily or optionally have to report data) which are related to

- speaker specific characteristics (e.g. sex, age, weight and height, smoking and drinking habits, socio-economic factors)
- regional/dialectal factors, and
- environment specific characteristics (environment of location of call, type of handset, type of network)

All the recordings will be performed over the telephone network through activation of a telephone server. Since the telephone interfaces are different from country to country there will not be a common recording tool used by all partners. Nevertheless, all recordings will be conducted via ISDN to simulate centralised services equipment. Data will be recorded on hard disk prior to creation of CD-ROMs. The speech data files description and representation is unified across languages. The SAM format is used as standard in SpeechDat. Annotation of the speech files is made at the orthographic level. In addition each database has a lexicon containing the standard orthographic words used in the annotations and corresponding canonical phonetic transcription using SAMPA notation.

2.4 Validation

Any database that is produced in the SpeechDat project must meet a set of minimum quality requirements in order to be approved by the consortium. The validation is carried out by SPEX. The validation criteria are those set by the predecessor project SpeechDat(M) and are tuned to the specification standards mentioned in section 2.2 above. To sum up, it will be examined if the proper items were recorded; if they are provided in sufficient quantities; if the speech files are properly annotated; if the database is well documented; if the concomitant lexicon is complete and well formated; and if the database in general adheres to the preset format specifications regarding directory structure and file names. A detailed account of the validation criteria is given in [11].

2.5 Distribution

It is obvious that the cost of developing resources like those of SpeechDat is prohibitive. Moreover spoken language resources have become fully commercial as the speech processing field has reached technical and commercial maturity. Therefore new channels and strategies for data distribution and commercialization are needed.

To address these problems, the European Language Resources Association (ELRA) was established as a non-profit organization in Luxembourg in February, 1995. The overall goal of ELRA is to provide a centralized organization for the collection, validation, and distribution of speech, text, and terminology resources and tools, and to promote their use within the European telematics R&D community. Additional activities include developing evaluation guidelines, serving as a broker between producers and users of LRs, and functioning as a central clearing house for information.

ELRA has started to publish a catalogue of existing databases, drafting a variety of viable contractual options (license agreements) for the suppliers and users of resources. The SpeechDat resources, once produced and validated, are transferred to ELRA for distribution. German [12], British English, French [13], and Swiss-French [14] databases produced in SpeechDat(M) are already on the catalogue. Sample files and documents are available at [1].

3 SCIENTIFIC AND COMMERCIAL RELEVANCE

Since the SpeechDat(M) project is finished and some databases are available (cf. section 2.5), first experiences exist as to the suitability of the databases for the rapid development of telephone-based speech recognition systems. In particular the application vocabulary, which is meant for immediate use in teleservices, has been tested. Different sites have conducted training and recognition experiments and reported word error rates between 1% and 2% for whole-word based HMM recognition of application words. If phoneme models, trained on the phonetically rich sentences, were used instead, the error rate increased by roughly a factor of two (cf. [14]). Similar results have been obtained for digit recognition.

In a different experiment the phonetically rich sentences of the German SpeechDat(M) database proved to be useful for training of balanced acoustic models for general German [15]. These models are necessary for recognition of new tasks that lack enough specific training data.

In addition to the acoustic training and testing material, the databases contain also important "language model" information, gathered from the spontaneous answers given to the questions asked during a recording session. As an example, answers to questions concerning a time response can be used to train a probabilistic bigram or finite-state network for time phrases.

While these first experiments already demonstrate the commercial usefulness of the SpeechDat(M) databases for the training of today's speech recognition teleservices, many more interesting scientific investigations can be carried out on the SpeechDat data, such as:

- Dependence of error rate on various factors such as: number of training speakers, age distribution, dialectal regions etc.
- Comparison of whole-word and subword unit based hidden Markov models
- Investigations about training/test mismatch between fixed and mobile telephone networks
- Performance benchmarking with respect to basic word error rate, keyword spotting performance and speaker verification performance
- Investigations of issues in porting speech technologies from one language to others evaluated on a multi-lingual language resource of so far unseen dimensions

4 FUTURE DEVELOPMENTS

Independently of the SpeechDat project currently a new consortium, called SpeechDat(E), is being formed. The aim of this consortium is to collect speech databases of East European languages including Russian which are conform to the SpeechDat specifications. There will be a big overlap between the members of the SpeechDat and the SpeechDat(E) consortium.

REFERENCES

[1] URL: http://www.phonetik.uni-muenchen.de/SpeechDat.html

[2] URL: http://www.spex.nl

[3] URL: http://www.icp.grenet.fr/ELRA/home.html

[4] R. Winski: *Definition of corpus, scripts and standards for Fixed Networks.* SpeechDat Technical Report SD1.1.1, 1996

[5] F. Senia et al.: *Environmental and speaker specific coverage for Fixed Networks.* SpeechDat Technical Report SD1.2.1, 1996

[6] J. van Velden: *Specification of speech data collection over Mobile Telephone Networks.* SpeechDat Technical Report SD1.1.2&SD1.2.2, 1996

[7] K. Kordi: *Definition of corpus, scripts and standards for speaker verification.* SpeechDat Technical Report SD1.1.3, 1996

[8] A. Nataf: *Environmental and speaker specific coverage for speaker verification.* SpeechDat Technical Report SD1.2.3, 1996

[9] F. Senia: *Specification of speech database interchange format.* SpeechDat Technical Report SD1.3.1, 1996

[10] F. Senia, J. van Velden: *Specification of orthographic transcription and lexicon conventions.* SpeechDat Technical Report SD1.3.2, 1996

[11] H. van den Heuvel: *Validation criteria.* SpeechDat Technical Report SD1.3.3, 1996

[12] C. Draxler: *The German SpeechDat Telephone Speech Corpus.* Proc. of SST 96, Adelaide, 1996

[13] D. Langmann, R. Haeb-Umbach, L. Boves, E. den Os: *FRESCO: The French Telephone Speech Data Collection - Part of the European SpeechDat(M) Project.* Proc. of ICSLP 96, Philadelphia, 1996

[14] A. Constantinescu, O. Bornet, G. Calloz and G. Chollet: *Validating Different Flexible Vocabulary. Approaches on the Swiss French PolyPhone and PolyVar Databases.* ICSLP 96, Philadelphia, USA, October 1996

[15] U. Bub, J. Köhler, B. Imperl: *In-Service Adaptation of Multilingual Hidden-Markov-Models.* ICASSP ´97