The SpeechDat(E) Project :

Creating Speech Databases for Eastern European Languages

Henk van den Heuvel (1), Valery Galounov (2) & Herbert Tropf (3)
(1) SPEX, University of Nijmegen, Nijmegen,Netherlands
(2) AudiTech-RD Ltd, Moscow, Russia
(3) Siemens AG, Corporate Technology Department, München, Germany
[H.v.d.Heuvel@let.kun.nl, auditech@neva.spb.ru, Herbert.Tropf@mchp.siemens.de]

Abstract

This paper gives some information regarding the SpeechDat(E) project for collecting telephone speech databases for teleservices in Eastern European countries. The project's position in the framework of other SpeechDat projects is explained; the organisation of the project is sketched; some details regarding contents and validation are presented; and finally, the current status of the project is briefly addressed.

Introduction

Recognition performance is the key to a successful teleservice. Satisfactory performance, however, is only achievable if realistic training data are available. The training data should comprise between several hundred and a few thousand speakers per gender depending on the number of speakers of the language in question. It should cover different dialects and accents and it should be representative of the telephone channel conditions likely to be encountered. Further, it is particularly important in Europe, that the service can be offered in different languages, asking for databases in various languages.

Within COCOSDA (an international scientific group supporting the creation and definition of spoken language resources) a standard for speech databases suitable for telephone applications was proposed (Polyphone Standard). Over that time, the Polyphone standard for telephony speech databases (fixed network recording) has been developed. The technical properties of Polyphone data sets are: 25 - 40 utterances per talker, both read and spontaneous; 5000 talkers; Telephone speech material collected digitally directly from the telephone network (a-law, mu-law) . See for a description of Polyphone standards Gibbon, Moore & Winski (1997).

The vocabulary of the database contains application-oriented words, digits; strings of digits, spelled words and names, sentences or phrases providing "phonetic coverage" of the language. The first Polyphone speech database was produced for American English by the American organisation LDC (Linguistic Data Consortium) funded by ARPA (Bernstein, Taussig & Godfrey, 1994).

So far, only a few telephone speech collections were carried out in Asia. A Japanese speech database ("Voice across Japan") and a still rudimentary Mandarin speech database collected in Taiwan are known to be roughly Polyphone-like.

In Europe, the projects SpeechDat(M) and SpeechDat(II) funded by CEC DGXIII/E represent the major industrial and academic players in Europe. Within SpeechDat(M) much of the groundwork of Polyphone-oriented standardisation was performed. In this case study the collection of 1000 speakers for each of 7 Western European languages was carried out. Based on this experience, currently in SpeechDat(II) 12 industrial and 5 academic partners are creating European telephone speech databases on a large scale:

- coverage of the 11 official languages of the European Union with their dialectal regions including the Swiss and Luxembourgish variants of French and German, and the additional languages Norwegian, Welsh and the Eastern European language Slovene, each database containing between 500 and 5000 speakers
- coverage of applications (application-oriented words, phonetically rich sentences)
- coverage of speaking styles (commands, carefully pronounced and spontaneous speech)
- coverage of environmental influences (mobile and fixed telephone network)
- suitable to develop and train robust speech recognisers and to develop and test robust speech verification systems for midterm applications

A summary of the SpeechDat(II) project and a brief evaluation is presented in Draxler, Van den Heuvel & Tropf (1998).

Other telephone speech databases were collected in Europe independently by Speech Expertise Centre (SPEX) and KPN Research in the Netherlands (Den Os et al. ,1995), by IDIAP in Switzerland (Constantinescu & Chollet, 1996), and by CSELT in Italy.

Based on the best experiences of SpeechDat(M) and SpeechDat(II) three other SpeechDat family members are now *in statu nascendi*: (a) SpeechDat-Car, aiming at the collection of at least nine databases for teleservices

control in car environments; (b) SpeechDat across Latin America, aiming at the collection of telephone speech databases in Middle and South America. The target is to collect at least seven databases for the Spanish and Portuguese language spoken in the regional variants of Latin America; (c) SpeechDat(E) aiming at the collection of at least three telephone speech databases for languages from Central and Eastern Europe.

The benefits of automatic speech recognition are likely to be especially significant in countries for which touch tone dialling is not yet widely supported and for which voice command provides the only advanced call management interface. Important benefits are likely to be gained in countries of Eastern Europe which have a developing infrastructure and economy, and in which increased modernisation and automation are significant factors for their economic development.

For Eastern Europe only two telephone speech databases are known, both currently being collected in very close relationship to SpeechDat(II) and exactly following SpeechDat standards: Slovene and Russian, each of them containing recordings of 1000 speakers.

SpeechDat(E) has the objective to extend existing and create new databases for languages in the middle and eastern part of Europe following the best practices and experiences of SpeechDat. Within SpeechDat(E) the Russian database will be extended to 5000 speakers, whereas for Czech and Slovak databases of 1000 speakers each will be created. The SpeechDat(E) consortium is still open for the other parties to join.

Project organization

The SpeechDat(E) project will be carried out within the COPERNICUS framework. Project duration will be two years starting in the summer of 1998.
At present the SpeechDat(E) consortium consists of three industrial contractors and three academic contractors. Siemens AG (Germany) acts as Project Coordinator, whereas AudiTech (Russia) acts as Scientific Coordinator.

The project focuses on the following databases:

- Russian (as spoken in Russia; 4000 recordings of different speakers; responsible: Auditech);
- Czech (as spoken in the Czech Republic; 1000 recordings of different speakers; responsible TU Brno);
- Slovak (as spoken in the Slovak Republic; 1000 recordings of different speakers; responsible: Slovak Academy of Sciences).

The project is organised in three work packages:
WP1 addresses the specifications of the databases in terms of database contents, speech annotation, data storage, speaker distribution, environmental coverage and validation criteria.
WP2 addresses platform installation, speaker recruitment, speech annotation, lexicon creation, and database documentation.
WP3 addresses the distribution of annotation tools and database validation.

Content and creation

The databases will closely follow the specifications defined in SpeechDat. The technical reports can be found via the SpeechDat(II) home page at:
http://www.phonetik.uni-muenchen.de/SpeechDat.html

All databases are recorded on telephone servers with ISDN connections. The signal format is 8bit 8KHz a-law, the European ISDN standard. For the annotation, the SAM file format has been chosen for two reasons: it separates signal from annotation data, and it is extensible. The annotations are encoded in ISO-8859, and a common SAM file format has been defined. The file system hierarchy is based on purely formal criteria, i.e. it is not content-related.
All SpeechDat databases can be addressed consistently in one large file system. File names follow the 8.3 character pattern of ISO-9660 for platform independence.
There is a large core content common to all SpeechDat databases. It consists of approximately 40 items that cover application words and phrases like digit strings, and phonetically rich words and sentences. An overview is presented in Table 1. The databases recorded in the SpeechDat(E) framework will cover more or less the same database items.

500+	4000+	Utterance description
2	**2**	**isolated digit items:**
1	1	single isolated digit
1	1	sequence of 10 isolated digits in one utterance
4	**4**	**digit/number strings:**
1	1	prompt sheet number (5+ digits, including any check digit)
1	1	telephone number (9-11 digits)
1	1	credit card number (14-16 digits, including any check digit) (a set of 150 numbers)
1	1	6-digit PIN code (a set of 150 codes)
1+	**1+**	**natural number**
1	**1**	**money amounts:**

1	1	currency amount, mixed size and units
0	0	large amount (e.g. main currency units)
0	0	small amount (e.g. including small currency units)
2	**2**	**yes/no questions:**
1	1	predominantly *yes* including 'fuzzy' yes/no (spontaneous)
1	1	predominantly *no* including 'fuzzy' yes/no (spontaneous)
0	0	approximately equal *yes* and *no* responses (spontaneous)
3+	**3+**	**dates:**
1	1	birthdate (spontaneous)
1+	1+	prompted date phrase, in word not digital format
1	1	relative and general date expression
2	**2**	**times:**
1	1	time of day (spontaneous)
1	1	prompted time phrase, in analogue not digital form
6	**3**	**application keywords/keyphrases**
1	**1**	**word spotting phrase using embedded application words**
5	**5**	**directory assistance names:**
1	1	city of birth/growing up (spontaneous)
1	1	most frequent cities (set of 500)
1	1	most frequent companies/agencies (set of 500)
1	1	proper name (forename and surname) (150 names)
1	1	proper name e.g. own forename (spontaneous) (set of 500+)
3	**3**	**spellings:**
1	1	real/artificial words to maximise letter coverage
1	1	spelling e.g. of directory assistance city name
1	1	spelling of proper name e.g. own forename (spontaneous)
4+	**4+**	**phonetically rich words**
9	**9**	**phonetically rich sentences**
43+	**40+**	**TOTAL utterances (excluding additional partner material)**

Table 1 - SpeechDat(II) corpus contents

The utterances will be annotated orthographically. Annotation is enriched with a set of markers for noises and deviations like mispronunciations and recording truncations.

Speaker recruitment is left to the individual partners, since the best strategy appears to be country-dependent. Still, some good strategies from the SpeechDat(II) project will be promoted, e.g. the strategy of appealling press releases in local new papers, and asking participants to recruit other participants in return for a small fee (snowball effect).

Each database comes with a lexicon containing one or more SAMPA phoneme transcriptions for each word in the database.

Validation

The SpeechDat(II) project is featured by a thorough validation protocol. The specifications which the databases should meet are evaluated by an independent validation centre. Validation proceeds in three steps:

1. Prevalidation of a small database of 10 speakers. The objective of this stage is to detect serious design errors before the actual recordings start.
2. Validation of complete databases. The database is checked against the SpeechDat specifications and a validation report is generated.
3. 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 a new report is written.

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

The validation of a database is carried out by the Speech Processing Expertise Centre, which was also in charge of database validation in the other SpeechDat projects and in the upcoming projects SALA and SpeechDat-Car.

A lot of validation work can be done automatically. For the SpeechDat(II) project software has been 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.

Current status

The Russian partner, AudiTech-RD, has collected a 1000 speaker database as an Invited (non-funded) Guest Partner of SpeechDat(II) in February of this year. It will be taken care of that there will be no speaker overlap with SpeechDat(E) recordings. This Russian database comprises 500 speakers from Moscow and 500 from St. Petersburg. The database was completed according to the specifications of the SpeechDat(II) project. Speech material (answers to 50 prompts) consists of spontaneous answers, reading digit sequences and text material (words and phrases). The total vocabulary is about 10,000 units. Recording was carried out through ISDN lines. The phoneme transcription for the lexicon was fulfilled

according to the Russian SAMPA table developed according to the requirements of St.Petersburg's phonetic school.
AudiTech has developed a number of programs to assist in SpeechDat database creation. The most important ones are LBO Expert, and prompt sheet maker.
LBO Expert is a transcription system elaborated especially for SpeechDat(II) and aimed at editing annotation files. In LBO Expert a transcriber selects a record, verifies if the speaker pronounced the text correctly and modifies the transcription if it is necessary. LBO Expert also allows to play waveforms using a standard Windows waveform output.
Prompt sheet maker is a non-interactive console application which takes text files with prompt sheet questions (question lists) as an input and produces a given number of prompt sheets in RTF (rich text format).

The SpeechDat(E) project was approved as Joint Research Project of the INCO-COPERNICUS Work Programme recently. The project will make it's official start in the summer of this year. Two more languages will probably be added to the database collection, depending on the question if appropriate funds can be raised.

References

Bernstein, J. Taussig, K. & Godfrey, J. (1994). Macrophone: an American English telephone speech corpus for the Polyphone project. *In Proceedings ICASSP-94, Adelaide, Australia,* Vol. I, pp. 81-84.

Constantinescu, D. & Chollet, G. (1996). Swiss Polyphone and PolyVar: building databases for speech recognition and speaker verification. In *Proceedings of the 3rd Slovenian-German and 2nd SDRV Workshop.*

Draxler, Chr., Van den Heuvel, H. & Tropf, H. (1998). SpeechDat experiences in creating multilingual speech databases for teleservices. In *Proceeding of the First International Conference on Language Resources and Evaluation, Granada, Spain.*

Gibbon, D., Moore, R. & Winski, R. (1997*). Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter. Berlin, New York.

Den Os, E., Boogaart, T., Boves, L. & Klabbers, E. (1995). The Dutch Polyphone Corpus. In *Proceedings of EUROSPEECH-95, Madrid, Spain,* Vol. I, pp. 825-828.

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