SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed

Henk van den Heuvel (1), Jerome Boudy (2), Zsolt Bakcsi (3), Jan Cernocky (4), Valery Galunov(5), Julia Kochanina (5), Wojciech Majewski (6), Petr Pollak (7), Milan Rusko (8), Jerzy Sadowski (6), Piotr Staroniewicz (6), Herbert S. Tropf (9)

(1) SPEX, A2RT, University of Nijmegen, the Netherlands
(2) Lernout & Hauspie France, Levallois-Perret, France
(3) Budapest University of Technology and Economics, Budapest, Hungary
(4) Brno University of Technology, Czech Republic
(5) AudiTech, Ltd, St.Petersburg, Russia
(6) ITA, Wroclaw University of Technology, Poland
(7) Czech Technical University in Prague, Czech Republic
(8) Institute of Control Theory and Robotics, Slovak Academy of Sciences, Bratislava, Slovakia
(9) Siemens AG, Munich, Germany

H.v.d.Heuvel@spex.nl

Abstract

In the Speechdat-E project five medium large telephone speech databases have been collected for Czech, Hungarian, Polish, Russian, and Slovak. The project was recently concluded. This paper reports briefly on the contents of the databases, elaborates on experiences gained from the data recordings and from the validation of the databases. The availability of the databases to the public is addressed, too.

1. Introduction

The main goal of the SPEECHDAT-E project - E stands for East - was the creation of multilingual spoken language resources (SLRs) to train voice-operated services for fixed telephone networks in central and Eastern European countries. In a multilingual environment as Europe, it is essential, that the end-user has access to 'common European Services' in his/her own native language and even dialect. SpeechDat-E extends the existing Western European language resources, which have already been recorded in the frame of Speechdat [1], with five major Eastern-European languages, viz. Russian, Czech, Slovak, Polish and Hungarian. The project started in December 1998 and was finalized in December 2000.

The SpeechDat-E SLRs include orthographical annotations and pronunciation lexicons. They are useful both for training and testing word models for typical present-day teleservices. In addition, phonetically rich sentences and words permit the training of more advanced, vocabulary-independent speech recognition systems on the basis of phone models. Table 1 gives an overview of all primary database features. These unique resources for Eastern-European languages have been realized by a consortium of 11 organizations.

These SLRs have been successfully validated by SPEX in Nijmegen, the Netherlands. Lernout & Hauspie (L&H) in Ieper and Paris had the responsibility of project coordination. L&H offered, together with Siemens, SPEX and Philips, their experiences established from earlier projects in speech data collection such as Speechdat [1] and Speechdat-Car [2]. ELRA (the European Language Resources Association) was associated to SpeechDat-E as funding partner for database validation, and as distribution center for the databases. With the exception of Hungarian, all SpeechDat-E databases are already publicly available via ELRA(http://www.icp.grenet.fr/ELRA/cata/tabspeech.html).

2. Design of the databases

A detailed account of corpus contents, and speaker and environment distributions is presented in [3]. Below follows a relevant excerpt.

2.1. Corpus items

The contents of all SLRs were derived from the item lists of SpeechDat. Table 2 summarizes the mandatory items in SpeechDat-E. Several databases contain optional, additional items. The corpora contain both read and spontaneous items. Due to the use of prompt sheets which were supplied in advance to participants, the items to be read were always known before the recordings.

2.2. Speakers

Proportional coverage of speakers with respect to speaker sex, age, and dialect was mandatory for each language. The sex distribution had to be 45-55% for each sex. At least 20% of the speakers should be 16-30 years of age, and another minimum of 20% between 31 and 45 years, and at least 15% between 46-60 years. The speaker accent distribution had to resemble that of the population with a minimum of 5% of the speakers in each accent region.

2.3. Recording environments

All SLRs were collected via the fixed network. The majority of the calls were realized from the environments 'home' and 'office'. A minority of calls came from public places like streets, booths, restaurants, etc. The proportion of calls from mobile phones was always below the permitted 5% limit.

Language	#Speakers	Owners	Producer	When available and how
Czech	1052	Lernout & Hauspie; VUT, CTU	VUT, CTU	Available at ELRA/ELDA (resource number S0094)
Hungarian	1000	Philips Speech Products, Aachen	Budapest University of Technology and Economics	Not available
Polish	1000	Siemens AG	ITA, Wroclaw Univ. of Technology	Available at ELRA/ELDA (resource number S0090)
Russian	2500	Siemens AG; Auditech	Auditech	Available at ELRA/ELDA (resource number S0099)
Slovak	1000	Slovak Academy of Sciences; Lernout & Hauspie	Slovak Academy of Sciences	Available at ELRA/ELDA (resource number S0095)

Table 1: The SpeechDat-E databases through a porthole.

No. of items	*Type*
2	isolated digits
4	digit / number string
1	natural number
2	money amount
2	Yes / no question
3	Date phrase
2	time phrase
6	application keyword (keyphrase)
1	word spotting phrase
6	directory assistance name
3	spelling word
4	phonetically rich word
12 (9)	phonetically rich sentence
48 (45)	**Total number of items**

Table 2: Summary of SpeechDat-E corpus items per prompt sheet. Numbers in parentheses are valid for the Russian SLR of 2500 speakers.

2.4. Orthographic transcriptions

All recorded data were annotated in SAM formatted label files according to the specification described in [4]. The annotations contain the orthographic transcription of the spoken utterances and additional information about speaker characteristics, recording conditions, and signal data file parameters.

The most important information in the label files are the orthographic transcriptions. These transcriptions are accompanied by additional marks, for:

1. Mispronunciations, signal truncations, and unintelligible parts,
2. Non-speech sounds from the speaker, e.g. hesitation, breath, lip smack, etc
3. Background environmental noise; short-time and long-time duration of the background noise are distinguished.

More details about orthographic transcription rules can be found in [6].

2.5. The lexicon

Each SLR contains a mandatory pronunciation lexicon of all words appearing in the database. It appeared that SAMPA standards of phonetic transcription were not defined for some languages; they had to be defined during the collection of these SLRs. Due to the high inflective character of Slavic languages, lexicons are usually relatively large. A typical example is the word *woman*: in English there are two different word forms (*woman, women*). In Czech there are ten different inflections and theoretically 14 (*žena, ženy, ženě, ženu, ženo, ženou, žen, ženám, ženách, ženami*).

All lexicons have the standard format defined for SpeechDat SLRs [4].

3. Experiences

This section addresses, per language, the experiences gained during the project, together with specific problems that had to be overcome. Before we address language specific experiences, we list a couple of general observations:

1. All partners used the same ISDN-controller (AVM: Fritz! or A1 card). Except for Hungarian, all recording platforms used the ADA (Automatic Database Acquisition) software interface. This software uses the CAPI 2.0 standard and was developed at UPC (Universitat Polytecnica de Catalunya) in Barcelona for the SpeechDat project. This combination of hardware and software yields very satisfying results.
2. The software for making the orthographical transcriptions differed per partner. All partners adapted in-house software to the new task. The many language-specific aspects of this task justify this heterogeneity in software.
3. Slavic languages are typically highly inflected. This has the consequence that less tokens of word-based items (e.g. application words, currencies) can be collected.
4. Successful speaker recruitment strategies differed per language. Most partners recruited representatives of (other) companies or private persons and made these responsible for speaker recruitment. Sometimes this was combined with a snowball strategy.

3.1. Czech

Minor problems were encountered during the creation of Czech database. These mostly pertained to the high number of inflections for Czech words. The items like natural numbers, money amounts, dates, times, etc. were designed with respect to optimal coverage of each particular inflection of the relevant words.

No problem appeared during the recordings. The quality of the landline connections seemed to be very good, since a minimum of calls were prematurely terminated (hang-ups usually originated from the speaker as a result of mistakes); the final average SNR of the recordings was also quite high.

Since both institutions involved in the recording were universities, mostly students were involved as recruiters. The advantage of the dual set-up was that it was easy to manage recordings over the whole country. Five Czech dialect regions were defined with the aid of a specialist from Masaryk University in Brno. The borders were placed at the borders of districts.

Special software was developed for the orthographic transcription (*FTP-transcriber)*. During the annotation also non-canonical pronunciations were marked (not required for SpeechDat-E). All transcriptions were checked for the syntax and possible spelling mistakes. In addition 5.2% of the transcriptions was hand-checked by listening tests.
The Czech SLR contains an additional transcription tier where pronunciations were marked. From this tier the lexicon could be automatically generated.

3.2. Hungarian

The Philips SpeechMania 2.2 dialogue manager program was used to record the calls. Connection quality was usually adequate, although in the beginning, the platform seemed to have problems finding appropriate recording start trigger levels. This lead to a number of calls, being discarded due to zero-length or truncated records.
The database contains a total of 1000 speakers for which the country was divided into four dialectal regions.
There was a subcontract with two country-wide companies (the Hungarian Railway Company and the Hungarian Telecommunication Company) for recruitment among their employees. There was only one contact person in each company, who was responsible for finding employees according to the parameters (age, sex and environment) that were specified on the prompt-sheet-covers. For the missing or unusable calls the prompt-sheets were reprinted and delivered to secondary schools, high schools and universities.
For transcription, A_TOOL was used (developed at BUTE-DTT). More information about this tool can be found at http://luna.ttt.bme.hu/speech/speechdt.htm. All annotators were native speakers; either experts from the university or students qualified on the field of speech.
In the first stage of the transcription process, completeness and quality of the session were checked. Incomplete and poor quality sessions were discarded. The second stage started with the entry of label file data fields, which couldn't be obtained automatically (calling region, accent, age, sex, environment, network). Finally, each item was orthographically transcribed by carefully listening to the recording.
Spot-checking each other's work helped the transcribers to improve consistency. After all recordings were annotated, a search was performed for various kinds of mistakes which were spotted during the annotation work, or which were expected to occur.

3.3. Polish

All recordings were made over the fixed Polish Network - TPSA.
In order to meet all demands (especially for phonetically rich material) special attention was paid to the proper generation of a set of 1280 different prompt sheets.
The recruitment of speakers was done in eight geographical-based regions into which the whole country was divided. In each region, depending on the population, a group of organizers was selected. Each organizer was responsible for providing full recordings of a group of 20 speakers from his/her region (each group had to be sex balanced and include people of predefined age ranges). The speakers were given the instructions, the sheets and were also briefed in person by the organizer. The recorded material was checked and in case of missing items or recording errors the organizer was asked for correction of the material. The chosen recruitment strategy gave good results.
Software used for the annotation of recorded material was prepared at the Institute of Telecommunications and Acoustics, at Wroclaw University of Technology, and specially tuned for the Polish database. The preliminary transcription of all spontaneous items and additional speaker information (such as: sex, age, accent etc.) were made in advance, viz. during the first check of correctness of recordings. Next, after typing in the proper sheet number, the annotator could listen to each speech file, correct the automatically generated orthographic transcription, and insert non-speech markers with special buttons.
The quality of landline connections was quite good (with high S/N ratio) and some minor low SNR signals were caused rather by environment of the recording place than by the connection.

3.4. Russian

The Russian speech database includes 2500 speakers. Four dialect regions were defined for the speaker selection.
We encountered some problems when making prompts for the items that include a lot of foreign words, i.e. city names and company/agency names. The problem was to compile large sets of the names (500 per set) whereas each of the names had to be both popular and 'readable' for the Russian speakers, i.e. preferably having just a single way of pronunciation. Sometimes foreign names provide possibility of accent variation, and this may cause hesitations. Nevertheless we could not avoid some important city names that remain difficult for Russians, e.g. *Reykjavik, Liechtenstein, Kuala Lumpur*.
Names of Russian companies were written in Cyrillic. For the names of foreign companies (e.g. *IBM, DHL, Panasonic*, etc.), both the Russian and the English versions were given in the prompt sheet, so that the speaker could choose the one that he/she found more convenient for reading. In the label files only the Russian equivalent was given.
The annotation procedure was carried out by native Russian experts by listening to the speech files with headsets in a quiet environment. Two programs were used for creating labels and orthographic transcription: LABEL and EXPERT. Both programs were developed by AudiTech specialists. LABEL is aimed at creating label files and can be easily used by non-professional staff. EXPERT is a transcription system developed for SpeechDat and upgraded for SpeechDat-E. The quality of each utterance was assessed after the transcription was completed.
The quality of speech signal transmitted through telephone channel in Russia varied widely. It may be qualified as 'clear', 'clear with background noise and outside signals', 'signal with strong background noise' and/or 'damaged signal' – down to the complete distortion. There were some cases of truncation of the recording session due to bad signal quality. This is caused by the Russian telephone network which is a mixture of digital and analogue switches.
Distribution of prompt lists by mail turned out as inappropriate. A lot of people refused to take part in the recordings regardless of detailed explanations of the aims of the research. Therefore, recruitment was done both on-site in the regions covered in the database and among the visitors from that regions in St. Petersburg. The most effective strategies were: A. The "snowball" method: the speakers were invited personally by our employee or by speakers who had already participated. B. Companies or other institutions were asked for help. The company manager helped engaging potential callers. The speakers were given the instruction

sheets and a brief explanation of the procedure (by our specialist) to avoid misunderstandings. Recruitment through the companies dealing with social research worked fine both in Moscow-St. Petersburg and in dialect regions.

3.5. Slovak

For Slovak three dialectal areas were distinguished.
In order to find suitable sets of phonetically rich words and sentences, a study was carried out on the probability of the Slovak phonemes in the spoken language [5]. A corpus of the texts from different areas (news, laws, ethnography, literature, poems etc.) was compiled. These texts were automatically transformed into an ortho-epic form using a software block of automatic orthographic to ortho-epic transcription formerly developed for a text to speech system. Thus, the preliminary statistic research on spoken Slovak was made without having any annotated Slovak speech corpus.
An oversampling method was used in prompt-sheet generation. It was regularly checked if the number of occurrences of every particular item fell in the acceptable interval. Some of the generated prompt-sheets were updated to contain the items missing.
In total, 2000 prompt sheets were generated, from which 1732 were distributed to the local recruiters. The number of calls was 1410, from which 1089 recordings were complete and 1000 were included in the final version of the database.
Annotation was done using the LABEL 1.0 software developed at the Slovak Academy. This tool included some basic syntax checking. It also monitored the balance of the sex, age and regional accent counts in the database so that the local recruiters could be instructed which speakers to recruit. It was easier to recruit female than male speakers.

4. Validation

Validation, as we will use the term here, refers to the quality evaluation of a database against a checklist of relevant criteria. These criteria are a compilation of the database specifications together with some tolerance margins for acceptable deviations of the specifications. General background information about SLR validation can be found in [8, 9]. A validation of all databases warrants that all databases adhere to the quality standards of the project and can therefore be exchanged within the consortium. In addition, the quality stamp of a successful validation is a positive signal to third parties when the databases become available through ELRA. The exact validation criteria for the SpeechDat-E databases are listed in [7]. They are basically the same as for SpeechDat. The most important deviations are the following [3]:

- Natural numbers may exceed 1 Million, provided they do not contain more than 4 significant digits;
- Phonetically rich sentences: each unique sentence should not appear more than 10 times;
- Phonetically rich words: each unique word should not appear more than 5 times;
- A table with the number of tokens of each phone at transcription level computed over the whole database is mandatory;
- Line lengths in label files may exceed 80 characters (extension of standard SAM format);
- The distribution of the dialect regions among the calls should be proportional to that of the population with a deviation of 5% at the maximum, and a minimum representation of 5% of the calls for each dialect region.

The approval of a database for the SpeechDat-E consortium was not decided by the validation center (SPEX), but by the project consortium on the basis of the validation report edited by SPEX. All databases were approved after the first validation. The quality of the databases is high in terms of completeness of recording sessions, documentation, consistency in database format, and transcription quality. The few remaining deviations from the validation criteria were only minor. The validation reports will be made publicly available through ELRA.

5. Acknowledgement

The SpeechDat-East project was supported by the EC in the INCO-Copernicus framework, code 977017.

6. References

[1] Höge, H., Draxler, C., Van den Heuvel, H., Johansen, F.T., Sanders, E., Tropf, H.S. “SpeechDat multilingual speech databases for teleservices: across the finish line”, *Proceedings EUROSPEECH'99*, Budapest, Vol. 6, pp. 2699-2702, 1999.

[2] Moreno, A.., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., Allen, J. “SPEECHDAT-CAR. A large speech database for automotive environments”, *Proceedings LREC 2000*, Athens, Greece, Vol. II, pp. 895 - 900, 2000.

[3] Pollak, P., et al. "SpeechDat(E) – Eastern European Telephone Speech Databases", in *the Proc. of XLDB 2000, Workshop on Very Large Telephone Speech Databases*, Athens, 2000.

[4] Pollak, P, Cernocky, J. *"Specification of Speech Database Intechange Format"*, SpeechDat-E Technical report ED1.3, 1999.

[5] Štefánik, J., Rusko, M., Považanec, D. “The Frequency of Words, Graphemes, Phones and other elements of the Slovak language” (in Slovak), *Jazykovedný Časopis*, Vol. 50, No. 2, pp. 81-93, 1999.

[6] Van den Heuvel, H. "Transcription Rules Applied to Speech Databases", SpeechDat-E Technical report ED1.4.1, 2000.

[7] Van den Heuvel, H.: *Validation criteria for SpeechDat(E) databases*, SpeechDat-E Technical report ED1.4.2, 1999.

[8] Van den Heuvel, H., Boves, L., Choukri, K., Goddijn, S., Sanders, E.P. “SLR Validation: present state of affairs and prospects”, *Proceedings LREC 2000, Athens, Greece*, Vol. I, pp. 435-440, 2000.

[9] Van den Heuvel, H. “The art of validation”, *ELRA Newsletter*, Vol. 5(4), pp. 4-6, 2000.

All SpeechDat-E technical reports referred to above with ED-codes can be accessed via: http://www.fee.vutbr.cz/SPEECHDAT-E/public/public_deliverables.html