MobilDat-SK – a Mobile Telephone Extension to the SpeechDat-E SK Telephone Speech Database in Slovak

Rusko Milan, Trnka Marian and Darjaa Sakhia

Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia
{milan.rusko, trnka, utrrsach}@savba.sk

Abstract

The paper describes design and process of collection, annotation and evaluation of a new Slovak mobile-telephone speech database MobilDat-SK, which is a mobile-telephone extension to the SpeechDat-E SK. The MobilDat-SK database contains recordings of 1100 speakers and it is balanced according to the age, accent, and sex of the speakers. Every speaker pronounced 50 files (either prompted or spontaneous) containing numbers, names, dates, money amounts, embedded command words, geographical names, phonetically balanced words, phonetically balanced sentences, Yes/No answers and one longer non-mandatory spontaneous utterance.
In the paper the structure of the database, the hardware and software solution of the automatic recording, the speaker recruitment strategy, the annotation process and evaluation process are described.
The MobilDat-SK database has been developed for the "Intelligent Speech Communication Interface" project in the frame of the State Research and Development Task "Building the Information Society".

1. Introduction

It is necessary for modern speech recognition systems to have large annotated speech databases available for training the acoustical models and testing the recognizer itself. As the development of such a database is a time consuming and expensive process and Slovak telecommunication market is relatively small, no professional quality Slovak speech databases had been built until the year 2000, when the SpeechDat-E initiative was started. With the growing importance of mobile telephone communications it became evident that a mobile telephone counterpart to the fixed line SpeechDat-E database has to be developed to be able to cover the specific features of the cellular telephone signal in the acoustic models and so to develop the recognizers for both GSM and PSTN networks.

2. SpeechDat databases

SpeechDat-E is a set of databases following the standard defined in SpeechDat II [1,2]. The collection was performed automatically via the ISDN telephone connection (on the recording side). As a compromise between the need and the economical possibilities, it was decided to build a 1000 speakers database for Czech, Polish, Slovak and Hungarian and a 2500 speakers database for Russian.

2.1. SpeechDat-E Slovak

After the preliminary statistical research a set of prompt sheets was generated. The prompt sheet is a list of sentences and words to be read by the caller and a set of questions to be answered. The prompt sheets were formed according to the possible areas of the speech recognizer applications (computers, banking, shopping, marketing, traveling and tourist information, telecommunication etc.).
Every of them included:

- isolated digits and their sequences
- digit / number strings
- natural number
- money amounts in Slovak crowns, Dollars and Euros and their smaller units
- yes/no questions (spontaneous answer)
- dates, prompted phrases with date, relative and general date expression
- time and time-phrases
- application words / key phrases
- word spotting phrase using embedded application word
- directory assistance names: city of birth (spontaneous), company, agency, surname, forename plus surname, own forename (spontaneous)
- spellings: artificial sequence, city name, own forename (spontaneous)
- phonetically rich words
- phonetically rich sentences

To reflex the real-life features the database was statistically balanced according to the:

- regional coverage - representation of the main phonetic groups. The repartition of speakers was proportional to the population in regions with 5% tolerance and with a minimum of 5% speakers per region.
- age of the callers
- sex of the callers

- acoustical environment

SpeechDat-E Slovak is the first large telephone speech corpus collected in Slovakia. It is available for users via ELDA.

3. MobilDat-SK database

The MobilDat-SK database for mobile telephone network was developed within the project "Intelligent Speech Communication Interface", in a frame of the State Research and Development Task "Building the Information Society". The structure of the MobilDat-SK database was designed to follow the SpeechDat specification [1,3] as closely as possible, and at the same time to serve as an extension to the SpeechDat-E Slovak database [4]. When used for speech recognizer training the combination of SpeechDat-E SK and MobilDat-SK databases should make it possible to produce robust speech models both for fixed and mobile telephone applications in teleservices.

The database was collected at the Slovak Academy of Sciences in Bratislava with close cooperation of the other partners - Technical University Košice (the coordinator of the project), Slovak Technical University Bratislava and Žilina University.

This database includes telephone recordings of 1100 speakers recorded directly over mobile telephones. The calls are routed to the PC based recording device connected to fixed PSTN using the digital ISDN interface.

3.1. Recording site and platform

The recording platform was located at the Slovak Academy of Sciences, Bratislava. It consisted of a PC connected to the ISDN line using the AVM FRITZ! card. This card supports COMMON-ISDN-CAPI 2.0. The recording software ADA was developed at UPC Barcelona [5]. The recorded data were daily transferred to the backup PC where the database had been collected and completed. Four computers connected to the backup computer with speech database were used for annotation and labeling of the database.

3.2. Speaker recruitment

The speakers were recruited from among the employees of the Slovak Academy of Sciences and teachers and students of the universities involved in the project as well as their relatives and their friends. It was checked regularly whether the recorded data followed the desired dialect, age and gender distributions. The most effective way of recruiting speakers was a personal contact with them via informed persons - recruiters - who knew how to explain them the need of recording the database. The recruiters were paid for the recruitment after the recordings of their callers were checked.

3.3. Design of prompting and prompt-sheet

Each speaker received a prompt sheet, i.e., a text with recording instructions and questionnaire for additional information on the speaker. The speakers could either complete and send the questionnaire via the internet, or send it to the Slovak Academy of Sciences via regular mail. The most important instructions for recording were repeated in the voice prompts of the automated recording system. For each speaker a unique prompt sheet was generated using predefined corpuses for each required item.

3.4. The database contents definition

We decided to use the database content from SpeechDat-E database, however according to assumed practical applications of the database we decided to extend it by some new items with one exception. The complete list of 50 mandatory items for MobilDat-Sk is shown in Table 1.

3.4.1. Additions to SpeechDat-E

We added 3 new items to the database to enrich the phonetical content:
O4 - sentence expressing question on departure or arrival of a train, including names of two train stations from the set of 500.
O6 - name of the town or tourism area from the set of 500 names.
In this item we tried to increase the number of realizations of the biggest towns included in SpeechDat-E SK (only two records of utterance of every town name are included there); the smaller towns were replaced by the tourism area names.
O9 - www or e-mail address from the set of 150 www and 150 e-mail addresses
R1 - One non-mandatory item – a longer spontaneous utterance was added at the end of the recording. The caller had to answer to a simple question from set of 25 like: "How to get from your house to a post-office?". This item should considerably extend the spontaneous content of the database.

3.4.2. Trimmings to SpeechDat-E

We decided to let out one item:
O5 - most frequent company/agency
The number of companies is too big to be covered reasonably; moreover many companies included in SpeechDat-E SK changed their names or does not exist any more. The use of this item seemed to be questionable.

3.5. Transcription

The transcription used in this database is an orthographic lexical transcription with a few details included that represent audible acoustic events (speech and non-speech ones) presented in the corresponding waveform files. The transcription is intended to be a rough guide that users of the database can further examine for details.

The transcriptions are made using the LABEL 2.0 transcription tool, developed at the Slovak Academy of Sciences. The screen of the program is divided into three sections: the prompt sheet, two edit fields, and series of buttons. The prompt sheet section contains all prompts of a sheet, the edit fields contain the complete prompt text and a field to enter the orthographic transcription. LABEL 2.0 performs various statistical analyses of the actual corpus. The actual count of annotated male and female speakers, people from each region, from age groups and other statistical values such as phonetical (phoneme) coverage can be calculated.

Table 1: 50 items which are mandatory for MobilDat-Sk.

Corpus identifier	Item identifier	Corpus contents	
A	1-6	6 application words	
B	1	1 sequence of 10 isolated digits	
C	1	1 sheet number (5 digits)	4 connected digit strings
C	2	1 telephone number (9-11 digits)	
C	3	1 credit card number (16 digits)	
C	4	1 PIN code (6 digits) (set of 150 SDB codes)	
D	1	1 spontaneous date (birthday)	3 dates
D	2	1 general date expression	
D	3	1 relative date expression	
E	1	1 word spotting phrase using an application word (embedded)	
I	1	1 isolated digit	
L	1	1 spontaneous spelling (speaker's forename)	3 spelled words
L	2	1 spelling of city name	
L	3	1 real/artificial spelling for coverage	
M	1	1 money amount in national currency	
M	2	1 international currency (dollar, euro) money amount	
N	1	1 natural number	
O	1	1 spontaneous name (own forename)	
O	2	1 place of birth (spontaneous)	
O	3	1 name of towns and cities (set of 500)	A new set chosen to enrich the SpeechDat set
O	4*	1 sentence expressing question on departure or arrival of a train (including names of two train stations from the set of 500)	Beyond the SpeechDat-E specification
O	6*	Name of the town or tourism area (set of 500 names)	Beyond the SpeechDat-E specification
O	7	1 "forename surname" (set of 150 SDB "full" names)	A new set chosen to enrich the SpeechDat set
O	8	1 "surname" (set of 150 SDB surnames)	A new set chosen to enrich the SpeechDat set
O	9	www or e-mail address (set of 150 www and 150 e-mail addresses)	Beyond the SpeechDat-E specification
Q	1	1 predominantly "yes" question	2 questions, including "fuzzy" yes/no
Q	2	1 predominantly "no" question	
R	1*	1 longer spontaneous utterance – answer on question	Beyond the SpeechDat-E specification
S	0-9	10 phonetically rich sentences	
Z	0-1	another 2 phonetically rich sentences	
T	1	1 time of day (spontaneous)	2 time phrases
T	2	1 time phrase (word style)	
W	1-4	4 phonetically rich words	

* *Only in MobilDat-Sk*

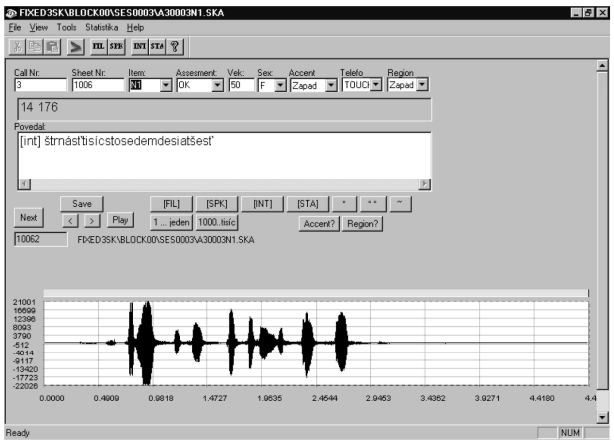

Figure 1: A typical display of Label 2.0 annotation tool

Annotators worked maximally four hours per day (resulting in about eight annotated speakers) to avoid errors due to tiredness. For the transcriptions, the ISO-8859-2 character set was used.

Transcription is case sensitive in the sense that all spelled characters (which have sometimes occurred in ”non/spelled items”, too) are transcribed in upper-case letters. All other texts are transcribed in lower-case letters.

No punctuation other than symbols used for special transcription purposes was used for the transcriptions. However, label files retain all punctuation provided to the speakers in the prompted text, including mistakes if these occurred.

Mispronounced nevertheless intelligible words are marked with one asterisk attached to the left of the word that is mispronounced, e.g. *transportation instead of the mispronounced “transportetation”.

Words preceded by asterisk include mispronunciations such as words with either extra or omitted syllables, asterisk is not used to indicate pronunciations of words that represent normal dialect or stylistic variants. In stretches of speech that are mispronounced, each mispronounced word is marked individually.

Words or stretches of speech that are completely unintelligible are denoted by a sequence of two asterisks: “**”. The “**” marker is separated from neighboring words by spaces.

Word fragments, i.e., instances in which the speaker did not complete a word, are considered a mispronunciation. It is marked accordingly with an asterisk attached to the left of the intended word. The full intended word follows, not a text fragment, as this could complicate the lexicon and create confusion if fragments were textually the same as valid words.

Words degraded by typical GSM noises and distortion are preceded by per cent sign: “%”.

Five categories of non-speech acoustic events are anotated.

[fil]: Filled pause. These sounds can be modeled well in a filled pause model in speech recognizers. Examples of filled pauses: uh, um, er, ah, mm.

[spk]: Speaker noise. All kinds of sounds and noises made by the calling speaker that are not part of the prompted text, e.g. lip smack, cough, grunt, throat clear, tongue click, loud breath, laugh, loud sigh.

[sta]: Stationary noise. This category contains background noise that is not intermittent and has a more or less stable amplitude spectrum. Examples: car noise, road noise, channel noise, GSM noise, voice babble (cocktail-party noise), public place background noise, street noise.

[int]: Intermittent noise. This category contains noises of an intermittent nature. These noises typically occur only once (like a door slam), or have pauses between them (like phone ringing), or change their color over time (like music). Examples: music, background speech, baby crying, phone ringing, door slam, door bell, paper rustle, cross talk.

3.6. Speaker demographic information, Accent/Region, sex, age.

There are nine mean dialect regions in Slovakia. It was decided to cover of them in our database; we controlled the coverage during the process of the database building. Nevertheless after the former discussions we found it appropriate to reduce the number of dialect regions by grouping them into 3 bigger groups for the final version of the database and for the evaluation.

Figure 2: Map of regions (west, center, and east).

Both sexes are represented by the same number of speakers. The age distribution of the speakers is shown in Table 2.

Table 2: Age distribution of the speakers

Age groups	*No. of speakers recorded*
under 16	32
16-30	558
31-45	281
46-60	196
over 60	33

3.7. Acoustical environment of the recordings

There are many possible environments from which mobile callers may conceivably make their call. In our case one speaker called only once, under one acoustic condition. So the required counts of the calls from different environments were specified as minimum 10 % of the database for every environment. The following 5 acoustic conditions have been chosen as representative in the mobile telephone communication environment:

Table 3: Acoustical environment

Acoustical environment	*Number of recordings in the database*
Home	547
Office environment	165
Public building (background talking)	149
Street (stationary pedestrian by road side)	121
Vehicle (passenger in moving car, train, bus, etc.)	118

3.8. Speech and label file format

Speech files are stored as sequences of 8-bit, 8-kHz A-law speech samples (CCITT G.711 recommendation). Each prompted utterance is stored in a separate file and each speech file has an accompanying ASCII label file in SAM format.

The MOBIL3SK\DOC\SAMPSTAT.TXT file contains a set of acoustic measures for each speech file in the database: maximal sample value, minimal sample value, number of samples, clipping rate, mean sample value, Signal-to-Noise Ratio.

3.9. The lexicon

The lexicon file is an alphabetically ordered list of distinct lexical items (essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions (ISO-8859-2). To keep the consistency the transcriptions use the same version of the Slovak SAMPA phonetic alphabet [6] as those in SpeechDat-E. List of used symbols is stored in the file SAMPALEX.PS. Entries are case sensitive.

3.10. Evaluation

The main purpose of the MobilDat-SK database is to enable research and development activities in Slovakia as well as a use for uducation purposes in the institutions which contributed to its development.

As the official evaluation by the SPEX or ELRA is expensive, the consortium has decided to cope with an internal evaluation among the members of the consortium. The evaluation process proceeded in three steps:

- In the first step a 10 callers database was built and evaluated
- The draft version of the 1100 callers database was made and passed to the consortium members for checking and tests in the second step
- After second-round annotation checking and making a new version of the lexicon with multiple pronunciations included and after taking some small recommendations of the consortium members into account, the final version of MobilDat-SK was finished and passed to the consortium for the final evaluation.

The database was accepted by the consortium.

4. FullDat SK, the combination of the two databases

Having two databases available – one for fixed network and one for mobile telephones, it is possible to compare results using acoustic models derived from the first one, the second one, or from a combination of both refered to as FullDat.

Such experiments were done by our colleagues from Košice Technical University. First they tested the quality of the SpeechDat-E SK database in comparison to SpeechDat II databases and also checked the possibility of crosslingual and bilingual recognition [7]. In the following experiment they compared the recognition rate when using models trained on all the three databases and test sets from both SpeechDat-E SK and MobilDat-SK. [8]

The Reference Recogniser REFREC training algorithm was chosen for the tests [9].

From all the experiments we present only the results obtained with tied_16_2 models (context dependent monophones with mixture of 16 Gaussians). Tests were made on the following item-groups: A – application words, I – isolated digits, Q – yes/no answers, O - directory assistance names, W - phonetically rich words and BC - connected digit strings.

Table 4: Comparison of word error rates (WER %) obtained when the recognizer was tested with the test set belonging to the same database from which the models were trained.

	A	**I**	**Q**	**O**	**W**	**BC**	**AVG**
SpeechDat	0,43	0,54	0.00	8,16	10,46	1,32	3,49
MobilDat	1,93	1,85	0.00	9,02	9,45	3,09	4,22

The recognition results are comparable, only slightly worse for MobilDat.

Table 5: Cross-tests. Comparison of word error rates (WER %) obtained when the recognizer was tested with the test set belonging to the different database than the one from which the models were trained.

train – test	**A**	**I**	**Q**	**O**	**W**	**BC**	**AVG**
SD – MD	1,70	5,09	0,00	20,24	23,76	5,57	9,39
MD – SD	1,37	2,15	0,00	11,28	20,03	2,03	6,14

According to the results, the recognizer with models trained on one database had considerably worse recognition rate when recognizing the test set from the second database and vice-versa.

Table 6: Comparison of word error rates (WER) obtained on the models trained on the entire databases with a test set that was a combination of both test sets. Grammar derived from FullDat was used for this test.

	A	I	Q	O	W	BC	AVG
SpeechDat	0,77	0,54	0,00	12,69	17,47	1,62	5,52
MobilDat	1,20	2,54	0,00	11,11	20,96	4,26	6,68
FullDat	0,99	1,57	0,00	11,63	17,91	2,94	5,84

The combination of the databases removes the influence of the diverse character of the speech data from the two telephone networks and the recognition results are comparable (only slightly worse) than those achieved with models trained on the entire databases with their appropriate test sets.

5. Discussion

According to the variety of acoustical conditions MobilDat SK contains much more background noises than SpeechDat-E. In such conditions the SpeechDat II noise annotation conventions seem to be insufficient. We plan to divide each of the non-speech acoustic events classes into several subclasses and to accomplish the re-annotation of noises.

The second change that we have started already is including alternative pronunciations into the lexicon.

One of the problems we do not know how to solve at the moment is an appropriate annotation of mail and web addresses. These are very often truncated or wrongly uttered by the speakers, as too many of them does not have enough experience with this kind of text. Similar problems were encountered with foreign words included in the databases, e.g. company names.

6. Conclusion

Liberalization of Slovak telecommunication market, hand in hand with recent boom in speech processing technology will hopefully lead to a competition among telephone operators and also other companies in the field of voice-driven teleservices. The created combination of databases can be the first step to professional design of such services in Slovakia and to bring them to practical life.

7. Announcement

This work was funded by the Ministry of Education of the Slovak Republic, task number 2003 SP 20 028 01 03 and by the Slovak Agency for Science, VEGA, grant No. 2/2087/22.

8. References

[1] Winski R., "Definition of corpus, scripts and standards for fixed networks", *Technical report, SpeechDat-II, Deliverable SD 1.1.1.*, workpackage WP1, January 1997, www.speechdat.org

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

[3] Pollak, P., Cernocky, J., Boudy, J., Choukri, K., Heuvel, H. van den, Vicsi, K., Virag, A., Siemund, R., Majewski, W., Sadowksi, J., Staroniewicz, P., Tropf, H., Kochanina, J., Ostrouchov, A., Rusko, M., Trnka, M. "SpeechDat(E) „Eastern European Telephone Speech Databases.", *Proceedings LREC'2000 Satellite workshop XLDB - Very large Telephone Speech Databases*, 29 May 2000, Athens, Greece, pp. 20-25.

[4] M. Rusko, "Definition of corpus, scripts, and standards for fixed networks," *Deliverable ED1.12.3, SpeechDat(E)*, 1999.

[5] F. Senia and I. Chatzi, "Instalation of the recording device and documentation," *Technical Report LE2-4001.SD2.1., Speechdat*, 1997.

[6] J. Štefánik, M. Rusko and D. Považanec, "Frekvencia slov, grafém, hlások a ďalších elementov slovenského jazyka," *Jazykovedný časopis*, Vol. 50, No. 2, 1999, pp. 81-93.

[7] Lihan,S.-Juhár,J.-Čižmár.A.: Crosslingual and Bilingual Speech Recognition with Slovak and Czech SpeechDat-E Databases. *Proceedingds of the 9th European Conference on Speech Communication and Technology Interspeech 2005*, Lisbon, Portugal, 2005, pp.225-228, (ISSN1018-4074)

[8] Lihan,S.-Juhár,J.-Čižmár.A.: Comparison of two Slovak speech databases in speech recognition tests. *33rd International Acoustical Conference ACOUSTICS High Tatras '06*, Strbske Pleso, Slovakia, October 4th - 6th, 2006 (accepted paper)

[9] Lindberg B., Johansen F.T., Warakagoda N., Lehtinen G., Kacic Z., Zgank A., Elenius K., Salvi G.: A noise robust multilingual reference recogniser based on SpeechDat(II), *Proceedings of the ICSLP 2000*, vol.3, pp. 370-373