ISSUES IN DESIGN AND COLLECTION OF LARGE TELEPHONE SPEECH CORPUS FOR SLOVENIAN LANGUAGE

Zdravko Kačič*, Bogomir Horvat*, Aleksandra Zögling†

*University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, 2000 Maribor, Slovenia
{kacic,bogo.horvat}@uni-mb.si

†University of Maribor, Research and Study Centre
Razlagova 22, 2000 Maribor, Slovenia
sandra.zogling@uni-mb.si

Abstract

In this paper, different issues in design, collection and evaluation of the large vocabulary telephone speech corpus of Slovenian language are discussed. The database is composed of three text corpora containing 1530 different sentences. It contains read speech of 82 speakers where each speaker read in average more than 200 sentences and 21 speakers read also the text passage of 90 sentences. The initial manual segmentation and labeling of speech material was performed. Based on this the automatic segmentation was carried out. The database should facilitate the development of speech recognition systems to be used in dictation tasks over the telephone. Until now the database was used mostly for isolated digit recognition tasks and word spotting.

1. Introduction

Design and collection of large telephone speech databases requires the compromise between large number of speakers with the limited amount of speech material per speaker or less speakers with much more speech material per speaker. Recently the SpeechDat project was accomplished with a number of databases recorded that contain large number of speakers (1000 or 5000 speakers) with approximate 5 minutes of speech material per speaker (Höge et al., 1997). The databases created achieved a good coverage of dialect, gender, and different call environments and surely represent a solid basis for development of various voice driven teleservices.

However, the lack of more speech data per speaker aggravates development and especially evaluation of systems in development phase for applications like dictation over the telephone. For development of applications like automatic telephone assistant with email reading and writing capabilities via voice, large vocabulary telephone speech databases with large telephone speech material per speaker would facilitate the development. Such databases could help especially in the pre-evaluation phase of the developed speech recognition systems.

2. Goals of the database design

The speech database SNABI was designed as a large corpus speech database to meet the development needs mentioned above. The requirements for the database was that it should contain on one side sufficient number of speakers to allow research in speaker independent speech recognition field and on the other to contain enough speech material per speaker to allow research in speech dictation task. The developed database should in this way facilitate development of speaker independent telephone speech recognition systems for wide range of applications including also speech dictation task. The database should therefore consists of different corpuses (isolated words, sentences from different domains, text passage) to meet these requirements.

2.1. Corpora of the database

The text corpus defined for the recordings was divided into a corpus of isolated words, number strings, and alphabet, general purpose text corpus (lingua), a task specific text corpus (MMC) as well as text passage. Table 1 shows an overview of the corpuses used.

Nr.	*CORPUS*	*CONTENT*
1	Words	80 words (command words, names of capital cities in Europe, digits)
2	Alphabet	25 letters
3	String	Number strings (set of 20 six digits long number strings)
4	MMC	Set of 529 different sentences - divided into three sub-corpora – MMC 1, MMC 2, MMC 3
5	Lingua	Set of 911 different sentences – divided into four sub-corpora – Lingua 1, Lingua 2, Lingua 3, Lingua 4
6	Passage	90 contextual dependent sentences

Table 1: Overview of the text corpora used in recording the database

The text corpus lingua consists of phonetically rich sentences and is aimed to cover wide range of speech telephone applications integrating also the dictation task. The MMC corpus was designed with the aim to cover different levels of spoken phenomena in particular constrained domain (office automation, railway information). The average length of sentences in the MMC corpus is 7.5 words, in corpus lingua 9 words and in the passage 15 words.

2.2. Database structure

The database contains speech and corresponding annotation files that are stored in a specified directory

structure. The directory structure is a content dependent directory structure, where files are organized according to the corpora used and further to sub-corpora. At the highest level the structure consists directories: words, lingua, and mmc. At the lowest level there are directories for corresponding sub-corpora (for example, lingua 1, lingua 2, lingua 3, and lingua 4, for the lingua corpus). Every directory at this level contains the speech material of all the speakers that read the text of the sub-corpus. The directory named /doc contains all the documentation of the database. The documentation consists of table of the SAMPA symbols used in phonetic transcription of the text, lexicon, speaker information table, and ISO88592 table.

Filenames follows the ISO 9660 file name conventions (8 plus 3 characters) according to the main CD-ROM standard. As it is useful for the user to be able to determine the content of the speech file by looking at the filename, the file name consists of different codes denoting speaker, recording type, corpus name, item ID, and file type. The following template is used:

XX Y V ZZZZ. NL M

where:
XX – denotes speaker ID code
Y – code of the sub-lexicon used
(W – words, N – numbers, A – alphabet, S – number string, M – MMC corpus, L – lingua corpus, P - passage)
V – type of recording (telephone)
ZZZZ – sentence ID number
N – type of pronunciation (W – isolated word, S – sentence)
L – language
M – file type (S – signal, W – segmentation and labeling on word level, P – segmentation and labeling on phone level)

For the signal and format the SAM format was used (Tomlinson et. al., 1988) and signal and data are therefore written in different files. Defining the structure of annotation file the SPEECHDAT format of annotation files was also considered (Draxler, 1998). In this way it is possible to update the annotation files with additional information not being limited with the header length as in case where a speech signal file has a header of fixed length. To each speech signal one or two annotation files are associated. Both files differ only in segmentation and labeling data. One file contains segmentation and labeling data for word level segmentation and annotation and the other for phoneme level. An example of an annotation file is given on Figure 1.

```
LHD: SAM, 5.10
DBN: SNABI_telephone
VOL: SI_tel_01
SES:
CMT: *** Speech file information ***
DIR: \SNABI\PHONE\WORDS\spk_0A
SRC: 0AWT1609.WSS
CCD: WORDS
CRP:
BEG: 0
END: 23694
REP: UMARIB, MARIBOR, SLOVENIA
RED: 12/Sep/1997
RET:
CMT: *** Speech data coding ***
SAM: 16000
SNB: 2
SBF:
SSB:
QNT: lin
CMT: *** Speaker information ***
SCD: OA
SEX: M
AGE: 45
ACC: 3
CMT: *** Recording conditions ***
REG: 2
ENV: OFFICE
NET: PSTN
PHM: ROTARY A
```

Figure 1. Sample annotation file

The annotation file shown on Fig. 1 is the same for word level and phoneme level phonetic transcription.

```
LBD:
CMT *** transcription data ***
LBR: ustavi
LBP: si u s t a: v i si
BSG:
0 23694 951 308 132 103 304 503 129 301
951 si u s t a: v i si
ESG:
```

Figure 2. Sample annotation file – part with word level transcription

Figure 2 shows an example of the annotation file for word level phonetic transcription, whereas Figure 3 shows an example for phoneme level transcription.

```
LBD:
CMT *** transcription data ***
LBR: ustavi
LBP: si u s t a: v i si
BSG:
0     6728  951         si
6728  7151  308         u
7151  8007  132         s
8007  9404  103         t
9404  12654 304 503     a:
12654 13776 129         v
13776 16869 301         i
16869 23694 951         si
ESG:
```

Figure 3. Sample annotation file – part with phoneme level transcription

Table 1 summarises the content of the speaker information table with additional information about speaker background.

Nr.	*Description*
1	Speaker ID code
2	Gender
3	Dialect regions: 1 – Štajersko

	2 – Koroško 3 – Gorenjsko 4 – Notranjsko 5 – Dolenjsko 6 – Primorsko 7 – Prekmursko 8 – Routarsko 9 – City of Maribor 10 – City of Ljubljana
4	List of lexica speaker read: Words Alphabet Strings (number strings) MMC: 1, 2, 3, LINGUA: 1, 2, 3, 4 Passage (90 contextual dependent sentences)
5	Education: 1- not finished elementary school, 2 – finished elementary school, 3 – vocational school, 4 – middle school, 5 – college, 6 – university, 7 – master degree, 8 – Ph.D.
6	Place of residence during the first years of elementary school
7	Place of residence during the longest period of life
8	Dialect region of parents (1 - 10)
9	Profession
10	Size in cm
11	Weight in kg
12	Reading activity: P – frequent, V – modest, R – seldom
13	Speech activity: Z – very active, A – active, M – moderate, R - seldom
14	Sickness (e.g., asthma)
15	Stimulants (e.g. cigarettes, coffee, alcohol)
16	Degree of funk during recording

Table 1: Speaker background information written in speaker information table

The lexicon is fairly simple as each entry consists of the orthographic form and of the phonetic transcription in the MRPA alphabet (MRPA, 1999).

3. Speaker selection and database recording

In contrast with the small vocabulary telephone speech databases that consists of hundreds or thousands of speakers, the selection of speakers in case of large vocabulary speech databases with several tenths of speakers can be more deliberate. In spite of that we did not defined any special criteria according to which the speaker selection should be made. The goal in speaker selection was to achieve good balance according to gender and age categories. The majority of speakers were students and employees of the university. Speakers were from different dialect regions although we did not cover all the main dialect regions in Slovenia. During the recording the speakers were not instructed to use standard pronunciation but rather to use their usual pronunciation.

All together 82 speakers were recorded. From these 31 were female and 51 male speakers. Majority of the speakers were around 21 years old, the youngest was 16 years and the oldest 62 years old. Six out of ten main dialect regions in Slovenia were covered by speakers.

All recordings were done in a normal office acoustic environment avoiding exaggerated presence of sounds like door slam, background music or cross talk, paper rustle, etc. The speakers read text from specially designed booklets with one item (word, letter, digit string or sentence) per page. In this way the speakers were able to concentrate only on the item uttered. They also did not try to hurry reading the text.

The speakers read different groups of sub-corpora making short pauses after each sub-corpus. The isolated words, alphabet, and set of digit strings were considered as sub-corpora during the recording.

Each speaker uttered in average 200 sentences, 80 isolated words, containing also digits, 20 digit strings and alphabet. From 82 speakers, 21 speakers uttered 450 sentences and 20 speakers also read the passage of 90 thematically connected sentences (simulating a voice dictation task). The database contains 21.416 recordings of sentences, 5760 isolated words and 5280 recordings of digits in digit strings. The total database consists of approx. 25 hours of telephone speech.

The database was recorded in a time span of 3 years. The calls were made over the analogue and digital telephone lines. The speech was first recorded with DAT recorder and then transferred over the digital connection (using DAT link) to workstation, where it was stored with 16 kHz sampling rate and 16 bit linear quantization with MSB-LSB byte order.

4. Segmentation

As the speech was originally recorded with DAT recorder the first segmentation was done on a word level during the data transfer to speech files on digital computer. The transfer and segmentation was done with proprietary software developed at the University of Maribor. During this process all the speech material was also manually checked.

For manual segmentation and labelling on phonemic level a proprietary software tool was used that was also developed at the University of Maribor. For labeling the MRPA alphabet for Slovenian language was used (MRPA, 1999). The developed tool allows positioning and inserting the segment borders. It further enables labeling of defined segments using the IPA symbols, which are transformed to appropriate MRPA symbols when the segments and labels are saved to a file. When the segments and labels are read from a file the MRPA symbols are converted to the corresponding IPA symbols, which are then displayed on the screen as labels. The tool also enables listening to individual segments or to an arbitrary selected segment of speech signal.

Part of the database was manually segmented and labelled at phonemic level. The rest of the database was automatic segmented with the HTK toolkit using the manually segmented speech material.

While developing system many different speech extraction methods have been tested. Results presented in this paper are obtained for the frame rate of 5 ms using 20 ms window lenght. Speech was analysed with mel-frequency cepstral coefficients, using also Δ and $\Delta\Delta$ coefficients and energy.

A simple 3 state left-right continuous density HMM models were used in which each observation probability distribution was represented by Gaussian density.

For segmentation evaluation purposes 300 randomly selected sentences were used for HMM training and a set of 400 randomly selected sentences were used for test. There was no overlap between the two sets.

The segmentation error was defined as a difference between automatically determined and manually placed segment boundaries.

The segmentation error can be quantitatively expressed by counting an automatic boundary positioning as correct if it falls within a given margin (the so-called "correct margin") of the reference segmentation. The number of correctly positioned boundaries divided by the total number of boundaries then gives the segmentation accuracy (Pauws, 1996). Figure 4 shows the preliminary results of speech segmentation accuracy.

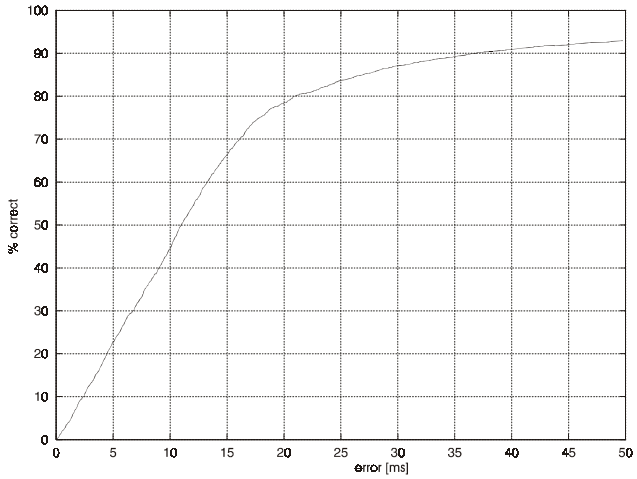

Figure 4: Preliminary results of database segmentation accuracy

Currently the work on increasing the segmentation accuracy is being performed.

An iterative procedure is planned where part of the automatic segmented material is manually verified and added to the existing manually checked material. This is then used as a basis for new iteration of the automatic segmentation procedure. The process will be stopped when the verification of automatic segmented material will show sufficient accuracy.

5. Database evaluation

The database was used till now in several experiments on automatic speech recognition. Mostly the tasks were isolated digit recognition and word spotting. In (Bub, 1997) the database was used in multilingual speech recognition experiment. In isolated digit recognition task reported in (Imperl, 1997) a comparative study of continuous density hidden Markov models and semi-continuous hidden Markov models was performed. In the implemented speech recognition system cepstral features were used, expandend with dynamic features (Δ and $\Delta\Delta$ coefficients) and diphone modeling was performed using the Laplacean probability density function. The experiments were performed for the SNABI and VoiceMail (for German language) speech databases. On the isolated digit recognition task, using two monolingual speech recognition systems with the same structure (for Slovenian and German languages), average recognition accuracy of 95.1% was obtained for the VoiceMail database and 98.6% for the SNABI database. The database was also used in the word spotting tasks (Kaiser, 1997, Kaiser 1997a) and in acoustic modeling with genetic algorithms (Kaiser, 1998).

6. Conclusions

The presented speech database is intended to facilitate development of telephone continuous speech recognition systems used in applications that would integrate the speech dictation task. The database was already used in various speech recognition tasks that mainly included isolated digit recognition and word spotting. The preliminary segmentation was performed and the iterative procedure with manual crosscheck of the automatic segmentation accuracy is foreseen to achieve higher segmentation reliability. In the future the evaluation of other parts of the database is foreseen, especially for continuous speech recognition and speech dictation tasks.

7. References

Bub, U., J., Köhler, B., Imperl 1997 *In-service adaptation for multilingual hidden Markov models*. In ICASSP´97, Munich.

Draxler, C., H. van den Heuvel, H. S. Tropf . (1998). *SpeechDat Experiences in Creating Large Multilingual Speech Databases for Teleservices*. LREC Proceedings, 1: 361-366.

Höge, H., H. S. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach & K. Choukri 1997. *European Speech Databases for Telephone applications*. In ICASSP´97, Munich.

Imperl, B., Z. Kačič, J. Köhler, 1997. *Isolated word recognition over the telephone using the Semi-continuous HMM.* In Proceed. Electrotechnical and Computer Science Conference.

Kaiser, J., 1997. *Word spotting in the telephone dialogue systems.* In Proceed. Advances in Speech Technology, Maribor.

Kaiser, J., Z., Kačič, 1997a. *Word spotting using phonetic fillers*. In Proceed. Electrotechnical and Computer Science Conference.

Kaiser, J., Z., Kačič, 1998. *Training of hidden Markov models with genetic algorithms*. In Proceed. Electrotechnical and Computer Science Conference.

MRPA, (1999), http/::www.phon.ucl.ac.uk/home/sampa/sloven-uni.html, SAMPA for Slovenian

Pauws, S., Y., Kamp, L., Willems, 1996. A hierarchical method of automatic speech segmentation for synthesis applications, *Speech Communication* 19 , p. 207-220

Tomlinson, M., R. Winski, W. Barry 1988. *Label file format proposal*. Esprit project 1542 (SAM): Extension Phase, Final Report.