SINOD - Slovenian non-native speech database

**Andrej Žgank, Darinka Verdonik,
Aleksandra Zögling Markuš, Zdravko Kačič**

Laboratory for Digital Signal Processing, University of Maribor
Smetanova ul. 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si http://www.dsplab.uni-mb.si

Abstract

This paper presents the SINOD database, which is the first Slovenian non-native speech database. It will be used to improve the performance of large vocabulary continuous speech recogniser for non-native speakers. The main quality impact is expected for acoustic models and recogniser's vocabulary. The SINOD database is designed as supplement to the Slovenian BNSI Broadcast News database. The same BN recommendations were used for both databases. Two interviews with non-native Slovenian speakers were incorporated in the set. Both non-native speakers were female, whereas the journalist was Slovenian native male speaker. The transcription approach applied in the production phase is presented. Different statistics and analyses of database are given in the paper.

1. Introduction

Slovenian language belongs to the group of highly inflectional languages with relatively free word order in a sentence. These set of language peculiarities makes Slovenian large vocabulary continuous speech recognition a very complex task. Although there is rather a large number of Slovenian speech databases' (Kaiser, 1998; Dreo, 1995; Kačič et al., 2000) considering the population size, it still exists a lack of speech resources for the development of Slovenian continuous speech recognisers. In a past few years, development of Slovenian Broadcast News language resources took place (Zögling Markuš et al., 2003; Žgank et al., 2004; Žibert., 2004). In addition to the development of speech resources also an emphasis was given on the development of Slovenian text resources (Žgank et al., 2005), which are needed for improved speech recognition language modeling for inflectional languages.

With the Slovenia's joining the European Union in 2004, an increased number of Slovenian non-native speakers were observed in different spoken media. Due to the characteristics of Slovenian language (inflectional morphology, free word order, dual,...) it is relatively difficult for a foreigner speaker to learn Slovenian literary language. This is consecutively reflected in the quality of spoken Slovenian by foreign speakers. There are several items influenced by this phenomenon. Some of them, important from automatic speech recognition point of view, are:

- *Phonemes from native language:* a non-native speaker uses instead of Slovenian phoneme the most acoustically similar phoneme from his mother tongue. Such borrowed phonemes should be taken into account by the speech recogniser's phonetic vocabulary (and partially by acoustic models).

- *Words' pronunciation:* some words in Slovenian language are mispronounced. Such words must be handled by acoustic models and vocabulary.

- *Words' order and grammar:* the order of words is technically correct, but unusual for literary language. This causes problems to statistical n-gram based language models, which are based on probabilities of word grams.

The analysis of BNSI Slovenian Broadcast News speech database showed a relatively small amount of non-native speech material in comparison to Broadcast News databases for other languages (Žgank et al., 2005; Pallett; Seymore et al., 1998). This was mainly caused by the time period, which was covered in the BNSI Slovenian Broadcast News project. The demand to improve a large vocabulary continuous speech recognition system for non-native speakers was the motivation to build the SINOD speech database - the first Slovenian non-native large vocabulary continuous speech database. One of non-native speakers was Russian; the other one was US-citizen. The database will be mainly used to adapt the Slovenian speech recogniser's acoustic models and phonetic vocabulary to non-native Slovenian speech. A part of SINOD speech database are also transcriptions, but probably the amount is to small to be able to use them to improve the speech recogniser's language model. Work on non-native databases for speech recognition in several other languages was reported until today: on English (Matsunaga et al., 2003; Fischer et al., 2003), German (Zhirong et al., 2003; Goronzy, 2003), French (Goronzy, 2003), Mandarin (Han et al., 2004),...

The SINOD speech database will be used as supplement to the BNSI Broadcast News speech database, as it has the same characteristics and structure (Žgank et al., 2004) as it was used in the BNSI project. The SINOD speech database consists from two interviews in the total length of 102 minutes with corresponding transcriptions. In the following sections, all major characteristics of the SINOD database will be presented.

The paper is organized as follows: the acquisition of raw material and the organizational aspects are described in Section 2. The observations gathered during the transcription work are collected in Section 3. The database analysis is presented in Section 4. The speech corpus description follows in Section 5 with short description of text corpus in Subsection 5.2. The conclusion and directives for future work are given in Section 6.

2. Acquisition of raw material

As it is hard to collect a large amount of spontaneous speech from non-native speakers for a language with relatively small number of speakers (Slovenian population size is approx. 2 million people), TV interviews, produced by Slovenian National Broadcaster RTV Slovenija, were investigated to find an appropriate material.

The raw material was acquired from the archive of RTV Slovenija in Ljubljana, which was our partner in the BNSI project. The original recordings of interviews were in analog format on Beta SP Master tapes. To preserve the quality during different subsequent processing stages the digital format, using DAT and DVD+R media, was employed for acquisition.

The audio signal, which is the main source for producing a speech database, was captured in linear raw format, using the 48 kHz sampling rate and 16-bit quantization. The audio signal in the SINOD speech database was then downsampled to 16 kHz and stored in the WAV format. The video signal was primary collected as assistance to transcribers, but it could be also used for the development of a demonstration application or to build a multi-modal system. The copy of raw video material was created on the DVD+R media. Later on it was grabbed onto personal computer and converted into MPEG-2 video files. Using the appropriate level of compression, such media format still enables good quality of video material, which could be used for any type of multi-modal applications.

As both shows were live interviews, scenarios from the Avid iNEWS system were not available. Broadcasts' scenarios were used as initial text for transcriptions (Žgank et al., 2005) by our transcribers in BNSI project. During the BNSI Broadcast News development process it was proven that such initial transcriptions speed up and ease the annotators work.

3. Transcription work

The annotation work can be very time consuming and tiresome. Therefore it is essential to organize annotator's workspace as user's friendly as possible, with easy access to all different sources that are needed. The most important tool, needed by an annotator is the Transcriber (Barras et al., 2001). It was developed by DGA and has evolved into an open source project, with support for many languages and wide area of usage in the speech technology community. In our project annotators used the Transcriber tool for the labeling and annotation of speech material.

A personal computer was the only piece of hardware needed during the transcription work. All other necessary resources for producing transcriptions were installed on it to ease and speed up the work. One of the most important additional resource was the interview's video, which was used as aid to transcribe difficult speech passages. Such passages were the parts with overlapping speech, hesitations, restarts,...

Annotator also had access to different vocabularies and newspaper text corpora to check particular word or word form. This was specially needed in the case of non-native speakers, as they used words infrequent in Slovenian language. They have also used some technical terms, when the topic of interview was connected with their profession. Access to already finalized BNSI Broadcast News transcriptions was made available to the annotator to assure the consequent level of quality in new transcriptions.

For transcription conventions, the Broadcast News guidelines defined by LDC (LDC homepage, 2006), LIMSI and COST 278 BN SIG (COST 278 BN SIG homepage, 2006) were used as baseline. They were first augmented with some language dependent rules already defined in the BNSI project (Žgank et al., 2004). Additional language dependent transcription rules were then added for the scope of Slovenian non-native speech database.

4. Database analysis

The transcription performed during the SINOD project was done by an annotator that had a lot of experience with transcription work in the Broadcast News domain. She had adopted the same three-stage approach that was used during the BNSI project (Žgank et al., 2005).

From speech recongiser's point of view it is essential to have acoustically homogeneous parts of speech, to be able to train high quality acoustic models. Consequently, the first part of transcription work was to segment the speech according to acoustical conditions (speaker, speaking style, channel, fidelity, background,...).

The speech signal was transcribed in the second step. This part of database creation process was very difficult, due to the spontaneous non-native speech in the interview. Several iterations were sometimes needed to complete a difficult passage. Some parts were labeled as unintelligible speech as it was impossible to produce unambiguous transcriptions.

The last step was used by the annotator to label different speech and non-speech events and to recheck again the difficult parts of transcription. This concluded the production phase of the first version of transcriptions. To improve the overall quality of speech database, additional supervision of transcription work was carried out.

The completed transcription was first spell checked by a linguist. She was also engaged in the preparation of transcription rules and helped solving all ambiguities that we were confronted with during the project. The annotator then corrected all errors in the first version of transcription. The transcription was once again checked after the spell checking by a second superviser that controlled other attributes of speech database. Here, a special emphasis was given to speaker's turns, speech and non-speech events and attributes of spoken material typical for spontaneous speech.

Approximately 32 hours of work were needed to finalize each interview. The amount of time needed for each interview was higher than for the news shows in the BNSI Broadcast News database. This increase was mainly caused by spontaneous non-native speaking style in the database.

The non-native spontaneous speaking style greatly differs from Slovenian literary language speaking style in the BNSI speech database. Non-native Slovenian speakers frequently use false words pronunciation, which can sometimes also partly (or even completely) change the meaning of a sentence. In the case, when speaker's native lan-

guage doesn't have an identical phoneme for a particular Slovenian phoneme, the non-native speaker uses acoustically best matching candidate from his native language. When a non-native speaker tries to recall the right Slovenian word he would like to use, he frequently uses false starts and hesitations that influenced the words' flow. In cases when he can't recall the correct Slovenian word, he uses the word from his mother's tongue and the journalist translates the word to Slovenian.
All such similar events were also observed in the non-native part of the Slovenian BNSI Broadcast News database (Žgank et al., 2005), although in smaller extent. The difference was mainly caused by the origin of non-native speakers in the BNSI database, as their mother tongue mostly belonged to south Slavic language subgroup and was as such similar to Slovenian language.

5. Speech and text corpus

In the following section all details of the SINOD speech database will be presented. Along with the statistics of the speech corpus, also details about the transcriptions will be given, as they may be useful for the adaptation of speech recogniser's vocabulary or even to improve the language model used during the decoding phase of non-native speech.

5.1. Speech database statistics

The SINOD speech database consists from two live interviews in a length of approx. 51 minutes each. The total amount of collected speech material in the SINOD speech database was 102 minutes. The topic of both interviews was general, mainly connected with the profession, work and past live of interviewee. The amount of collected speech material is approximately 4.72% of the complete BNSI Broadcast News speech database and approximately 5.67% of the BNSI acoustic models training set (Žgank et al., 2005).
Only two speakers were present during each interview. A native Slovenian male journalist interviewed a female non-native speaker. In the first interview the interviewee was of Russian origin. The level of non-nativeness in her speech was relatively low as she uses the Slovenian language in her profession for more than two decades. It must be also taken into account that Russian and Slovenian belongs to the same Slavic language group and are as such related. The most frequent non-native event was wrong pronunciation or slightly unusual word order. Another interesting non-native event was the usage of Russian words that exist in the same Slovenian form, but have partially or completely different meaning.
The interviewee in the second broadcast was an US-citizen. She lives in Slovenia for a last few years and had contact with Slovenians for less than a decade. The level of non-nativeness in her speech was relatively high, especially in the parts with emotional speech. Along with other non-native effects, she had problems to recall some Slovenian words, especially those in dual word form. Also the word order in sentences was sometimes unusual for Slovenian literary language.

To be able to denote different acoustic conditions of a speech database, different focus conditions (f-condition) were defined for the Broadcast News databases (Schwartz et al., 1997):

- *F0:* read speech, studio
- *F1:* spontaneous speech, studio
- *F2:* telephone,
- *F3:* music in background
- *F4:* presence of noise
- *F5:* non-native speech
- *FX:* mixed conditions

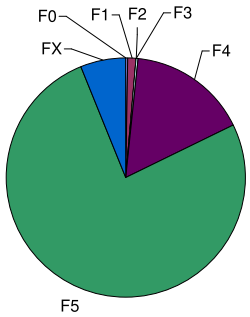

Figure 1: F-condition ratios for the SINOD database.

The analysis of f-conditions is given on Figure 1. The majority of speech (76.2%) in the SINOD database belongs to the focus condition F5, where non-native speech is placed. The next biggest portion of speech (16.2%) is labeled as speech with presence of noise - condition F4. The only other representative group is condition FX with 6.0% of speech. Here, speech with mixed attributes can be found. Other three conditions (F0, F1, F3) have a minimal portion of speech. There was no speech with low-fidelity (F2). Such distribution of f-conditions is non-representative for a typical Broadcast News database, where the majority of speech usually belongs to conditions F0 and F1. There are two main causes for this alteration in distribution: the first one is the presence of non-native interviewee, the second one is the type of broadcast (interview).
Some additional statistics for both broadcasts are given in Table 1.
Altogether, there are 2044 dialog turns in SINOD database, where the average acoustically homogenous turn last for 2.9 seconds. The duration of longest turn was 11.9 seconds. As SINOD database belongs to non-native speech resources, it is of particular interest to investigate words that reflect non-native nature: 654 words were mispronounced and 305 were truncated. This represents 7.7% of the whole speech corpus.

	SINOD
Number of turns	2044
Average turn length (s)	2.9
Longest turn (s)	11.3
Shortest turn (s)	0.2
Mispronounced words	654
Truncated words	305

Table 1: SINOD speech database statistics.

5.2. Text corpus

Transcriptions of both interviews from the SINOD speech database consist from 12.5k words, where 2516 of them were different. The BNSI speech corpus contains 268k words, from which 37k are different words (Žgank et al., 2005). The overlap between SINOD and BNSI vocabulary was 69.9%, which reflects the inflectional nature of Slovenian language that influences the performance of a large vocabulary continuous speech recognition system.

The amount of text material produced in the SINOD database project is probably too small to be able to adapt speech recogniser's language model to a non-native speaking style. But it still can be used to analyse events in spoken non-native Slovenian that influence the language modeling. Here a particular interest could be given to order of words in a sentence.

6. Conclusion

The SINOD Slovenian non-native speech recognition database will be used in the development process of a Broadcast News transcription system. In such a way the modelling of non-native Slovenian speakers will be improved. The possible area of usage is also multilingual and crosslingual acoustic modelling, as non-native speaker usually share phonemes between two or even more languages.

Acknowledgment

Authors would like to thank the staff of RTV Slovenia that helped us to collect all the necessary data. Special acknowledgment goes to the annotator who had transcribed the SINOD speech database.

7. References

Barras, C., Geoffrois, E., Wu, Z. and Liberman, M., 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, Vol. 33, Issues 1-2, 5-22.

COST 278 Broadcast News SIG homepage http://cost278.org/bn

Dreo, D., 1995. Slovene speech data base SNABI. *Dialog Man-Machine : second International Workshop*, Maribor, Slovenia.

Fischer V., Janke, E., Kunzmann, S., Ross, T., 2001. Multilingual acoustic models for the recognition of non-native speech. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 01*, Madonna di Campiglio, Italy.

Goronzy, S., Eisele, K., 2003. Automatic pronunciation modelling for multiple non-native accents. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 03*, St. Thomas, U.S. Virgin Islands.

Han Y., Yuanyuan P., Hong W., Zhengpeng, Z., 2004. An acoustic-phonetic analysis of large vocabulary continuous Mandarin speech recognition for non-native speakers. *International Symposium on Chinese Spoken Language Processing 2004.*, Hong Kong.

Kačič, Z., Horvat, B., Zögling A., 2000. Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language. *Proc. Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Kaiser, J., Z. Kačič, 1998. Development of the Slovenian SpeechDat database. *Proc. Speech Database Development for Central and Eastern European Languages*, Granada, Spain.

LDC homepage http://www.ldc.upenn.edu

Matsunaga, S., Ogawa, A., Yamaguchi, Y., Imamura, A., 2003. Non-native English speech recognition using bilingual English lexicon and acoustic models. *Proc. ICASSP 2003*, Hong Kong.

Pallett, D. S., 2002. The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program. *Speech Communication*, Vol. 37, Issues 1-2, 1:3–14.

Schwartz, R., Jin, H., Kubala, F., Matsoukas, S., 1997. Modeling those F-Conditions - or not. *Proc. DARPA Speech Recognition Workshop*, pp 115-119, Chantilly, VA.

Seymore, K., Chen, S., Doh, S., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M., Stern, R., Thayer, E., 1998. The 1997 CMU Sphinx-3 English Broadcast News Transcription System. *Proceedings of the 1998 DARPA Speech Recognition Workshop*.

Zhirong W., Schultz, T., Waibel, A., 2003. Comparison of acoustic model adaptation techniques on non-native speech. *Proc. ICASSP 2003*, Hong Kong.

Zögling Markuš, A., Žgank, A., Rotovnik, T., Sepesy Maučec, M., Vlaj, D., Hozjan, V., and Kotnik, B., 2003. Spoken Language Resources at University of Maribor. *Proc. of 10th International Workshop Advances in Speech Technology 2003*, Maribor, Slovenia.

Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., Kačič, Z., Horvat, B 2004. Acquisition and annotation of Slovenian broadcast news database. *Fourth international conference on language resources and evaluation, LREC 2004*, Lisbon, Portugal.

Žgank, A., Verdonik, D., Zögling Markuš, A., Kačič, 2005. BNSI Slovenian Broadcast News database - speech and text corpus. *Proc. Interspeech 2005*, Lisbon, Portugal.

Žibert, J, Mihelič, F., 2004. Development of Slovenian broadcast news speech database. *Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.