Acquisition and Annotation of Slovenian Broadcast News Database

Andrej Žgank, Tomaž Rotovnik, Mirjam Sepesy Maučec, Darinka Verdonik, Janez Kitak, Damjan Vlaj, Vladimir Hozjan, Zdravko Kačič, Bogomir Horvat

Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova ul. 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si http://www.elektronika.uni-mb.si

Abstract

This paper presents the Slovenian Broadcast News Database project that was started in year 2002 as cooperation between University of Maribor and Slovenian national broadcaster RTV Slovenia. The resulting database will be used for large vocabulary continuous speech recognition and multimedia database retrieval or archive indexation. First some organizational aspects that were needed in initial phase of the project are described. The raw audio and video material was acquired from the original Analog Beta SP Master tapes that are preserved in the RTV Slovenia's archive. Raw material was copied to DAT and DVD media. Also additional teletext material was collected. The manual annotation of speech material is performed with the Transcriber tool. The annotation rules were defined on the basis of general rules for Broadcast News databases, with some special language dependent sections. Also some statistics on a part of current material are given.

1. Introduction

The speech recognition research area is confronted with broad spectrum of tasks, where continuous speech recognition with large vocabulary is being one of the most challenging. One of the main research efforts in this area is being the Broadcast News speech recognition that was launched in the year 1996 by NIST and DARPA (Pallett, 2002). The idea behind Broadcast News is to use "found" speech instead of specially recorded speech and then develop continuous speech recognisers. Those can be used for indexing TV material, topic detection and tracking, subtitling, etc. Several such systems are already used in commercial applications (Cole et al., 1995; Imai et al., 2000; Brousseau et al., 2003). Broadcast News speech databases were created for different languages, amongst other for: American English, Mandarin, Spanish, Italian, Czech, etc.

Slovenian language is one of the smallest official languages in Europe with approximately 2 million speakers. The first large Slovenian speech database SNABI suitable for speech recognition was built in 1995 at University of Maribor (Dreo, 1995). As the speech recognition research area evolved, different new Slovenian speech databases were created. Some of them are: SpeechDat(II), PoliDat, Interface, Gopolis, ... Slovenian language belongs to the group of inflectional languages, which makes large vocabulary continuous speech recognition an extremely hard task. According to analysis (Sepesy Maučec, 2001) ten time larger Slovenian recognition vocabulary is need to assure the same text coverage as in English language. This was one of the main reasons that until now no Slovenian speech database was built for work on general topic. The only two Slovenian speech databases suitable for continuous speech recognition were SNABI (Kačič et al., 2000) and Vreme (Martinčič-Ipšič et al., 2002). Database SNABI covered some general topics but the other part of it is limited to traffic information systems. Database Vreme is limited to domain of weather forecasts. Considering these facts, the development of Broadcast News database for Slovenian language was initiated with the intention to create first Slovenian speech database for general domain.

The Slovenian Broadcast News Database (Zögling et al., 2003) project was started in the last quarter of year 2002 when the project's contract was signed. The partners in the project are University of Maribor, Slovenia which has strong background in speech database development process, and Slovenian national broadcaster RTV Slovenia. The goal of the project is to develop complete Broadcast News speech database for Slovenian language which will be also available through ELRA/ELDA. The Slovenian Broadcast News database will be used for different research purposes in area of large vocabulary continuous speech recognition and multimedia database retrieval or archive indexation.

The paper is organized as follows: organizational aspects considered during the preparation phase are described in Section 2. The raw TV material collecting procedure is presented in Section 3. The ongoing work on annotation and some statistics can be found in Section 4 and 5. Conclusion and description of future work is given in Section 6.

2. Organizational aspects

In general, there are two possibilities of collecting speech material for Broadcast News system. The first one is to record the material from broadcast transmission on air or cable. In this case only current shows can be collected. The disadvantage is that transmission channel can degrade the quality of signal. The second approach is to copy the speech material from archive. In this case the quality is usually better and also access to older material is granted, which simplifies the collection material over longer period. Due to cooperation with Slovenian national broadcaster RTV Slovenia the second approach was used in the project.

During the organizational phase analysis of existing Broadcast News databases was carried out. The main characteristics taken into account were:

- type of shows - news, talk shows, interviews, ...

- quality of speech - amount of different ”F” classes,
- speakers’ characteristics - number, gender, diversity,
- topics’ diversity over time span.

According to results of this analysis raw structure of new speech database was created. TV shows that should be collected in the database were selected from TV timetable. The time span of 4 years (between 1999 and 2003) was set for the coverage of TV shows. An analysis of TV programme for this time span showed that proper representation of female speakers in interviews and talk shows would be very hard to achieve, so some of the TV shows selected for the database were also elder. The intention was to collect 72 hours of raw broadcasted material. The final Slovenian Broadcast News database should have the following structure:

- Training set: 30 hours of speech material, used for acoustic training
- Development set: 3 hours of speech material, used for tuning the parameters,
- Test set: 3 hours of speech material, used for evaluation.

3. Acquisition of raw TV material

The TV shows stored in the archive of the RTV Slovenia were originally recorded on Analog Beta SP Master tapes. One part of shows used only mono audio channel and the other part two channel stereo. The video signal was in PAL format. To acquire the raw material Analog Beta player was connected to the DAT audio recorder and DVD+R(W) video recorder. The audio material acquired on DAT media was then used as main source for creating the speech database. The DVD+R media version of video and audio was needed in annotation procedure as additional source of information. This was mainly used to identify the speakers and to determine the acoustic conditions, but it can be also used in future for some image processing research. The recording conditions for both types of involved media are presented in Table 1.

Media type	Stream	Recording conditions
DAT	Audio	48 kHz sampling rate 16 bit resolution 2 channels
DVD+R	Audio	48 kHz sampling rate 16 bit resolution 2 channels 256 kbps bitrate AC3 compression
DVD+R	Video	5.1 Mbps bitrate MPEG2 compression

Table 1: Audio and video recording conditions on DAT and DVD+R media.

Audio signal was recorded on DAT uncompressed in raw format. On the other hand, AC3 audio compression with 256 kbps was used for audio signal on DVD+R media. The video signal was digitalized and compressed with MPEG2 codec. The compressed video bitrate was 5.1 Mbps.

Although the material collected on DVD+R media is only intended to be used as help during the annotation and maybe in later stages of the project for some image processing, the frequency analysis comparison between DAT audio and DVD audio was carried out to analyze the influence of AC3 audio compression. The result for a segment of speech signal is presented in Figure 1.

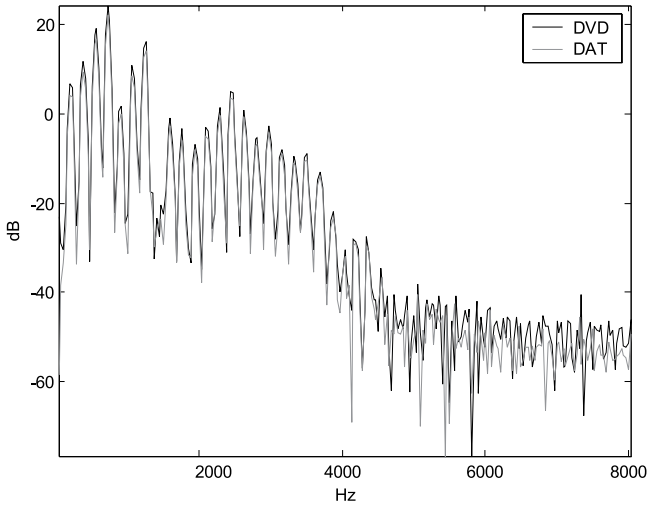

Figure 1: Frequency analysis of speech segment captured from DAT and DVD+R media.

Figure 1 presents result of frequency analysis. The speech segment used is a part of TV show, where the speaker makes a field report over the telephone line - the bandwidth of speech signal is limited to 3.4 kHz. Speech signals from both media types have almost identical frequency characteristic, although the DVD signal was compressed with AC3 codec. At higher frequencies there is some small difference between both frequency characteristics, but this divergence shouldn’t have significant influence on speech recognition performance (Barras et al., 2001/1) if video and audio stream from DVD+R media would be used in a multi-modal application.

Audio signal frequency analysis was also performed before each show was copied. With this additional step the quality of raw material was assured. Altogether 95 different TV shows were included in the raw material, some of them are:

- evening news,
- late night news,
- talk shows,
- interviews,
- magazines.

As Slovenian language belongs to the group of inflectional languages, special care was also given to the raw text material needed for language modeling. Different text

material in form of teletext subtitles was collected for this purpose. The smaller part of subtitles was used during the annotation procedure as initialization material.

The rest of the teletext subtitles will be used as separate text corpus for generation of language models. Such separate text corpus can be used for interpolation with normal text corpus collected from newspapers to decrease the difference between both styles of text.

4. Annotation of speech material

As Broadcast News type of speech databases can be seen as a standard, the transcription conventions created by LDC (LDC Homepage, 2004) where taken as a baseline. These rules were supplemented with transcription directions produced by LIMSI (LIMSI Homepage, 2004) and the document generated on the COST 278 Broadcast News Workshop. The following transcription rules were summarized from the above mentioned sources.

4.1. Some language dependent rules for orthographic transcription

4.1.1. Acronyms

Sign tilde (~) was used to mark spelled pronunciation of acronym (~ZDA), sign at (@) to mark normal pronunciation of acronym (@NATO). Declination of acronyms can be marked with hyphen (@SAZU-ja, ~APZ-ja, ~BBC-ja) or by changing last letter of acronym according to declination endings (@NATO @NATA, @UNESCO @UNESCA), it depends on type of declination.

4.1.2. Capitals

Capital letters were used only for acronyms (~APZ, ~NBA, @SAZU), spelling (~S ~I ~M ~O ~N ~I ~T ~I) and proper names (Moskva, New York, George ~W. Bush). Common words used as proper names were decapitalized and written between quotation marks (”ministrstvo za notranje zadeve republike Slovenije”, ”športno društvo Smolnik”, ”združene države Amerike”).

4.1.3. Punctuation marks

Punctuation marks were used to indicate syntactic and prosodic boundaries. We’ve used the following:

- . end of a sentence
- , separating clauses, pauses in speech etc.
- ? question mark
- ! exclamation mark
- : colon
- ... unfinished clause

Space was put before and after punctuation mark (dela , dela , vsako leto bolj !).

4.1.4. Compound words

Compound words were written according to Slovenian orthography (rdeče-bel, anglo-ameriški) (Toporišič et al., 2001). The exceptions are compound words including digit or acronyms. They are written without hyphen and with tilde or at sign, depends on pronunciation (for example ~TV spored, ~C vitamin, ~CD @ROM).

4.1.5. Abbreviations

Some abbreviations were allowed during the transcription to ease the work. Those were: itd., ipd., tj., OK, npr., oz. Abbreviations of academic titles (like dr., prof.) were not allowed, since they can be declined (doktor - doktorja). The same holds for numbers and dates - they are always written with whole word (dvajset celih pet, stoosemintrideseti, (Ludvik) štirinajsti).

4.1.6. Internet

Web addresses and e-mail addresses are written as usual (www.google.com; xy@yahoo.com) and marked with Edit/Insert event/Pronounce/...

4.1.7. Foreign words

There’s no special treatment of foreign words which are part of Slovenian vocabulary (like judo, premier). All other words in foreign languages are marked with Edit/Insert event/Language or with Segmentation/Edit section attributes/nontrans, when the segment of words in foreign language is long. Foreign proper names are written as usually in Slovenian orthography (George Bush, Tony Blair).

4.2. Other transcription rules

The main intention is to use the Slovenian Broadcast News database for speech recognition. Therefore each speech segment should have homogenous acoustic conditions. The acoustic quality attributes presented in Table 2 were used.

Fidelity	Full bandwidth	Limited bandwidth
High	Studio	Sounds clear
Medium	Field	Noisy
Low	Channel noise	Not intelligible

Table 2: Acoustic quality attributes used during annotation of speech material.

The complete show was divided in to section blocks that also represent homogenous unit. The most frequent block is a report that carries the story. Other blocks are fillers, which denotes headlines or story announcements and non-transcribed blocks, which covers jingles and commercials. Each segment block is then segmented into smaller parts - speaker turns. The breakpoint is inserted at any possible event to keep the manageable length for speech recognition.

To standardize the set of tags, a close list of them was set up. All transcribers should use this list to keep the same tags for same acoustic event. Mainly non-lexeme words and interjections fall in this category.

The software tool used for annotation is the Transcriber (Barras et al., 2001/2). Because video was also captured in digital format the computer is the only tool needed during the annotation procedure, as the speech material was transfered from DAT to PC. The usage of TV and video-recorder is obsolete. The video material necessary during the work (to identify speakers, to analyze difficult part,...) can be played on computer via the DVD-ROM. For a part of speech material text draft was used as initial material, but

it was found out that the annotation procedure was not simplified in comparison to the case where no initial material was used.

5. Current status and statistics

It is expected to finish the annotation phase of the project at the end of 2004. On a part of annotated speech material some statistics were collected to represent the variety of speech material - Table 3.

	Show 1	Show 2	Show 3
Length (s)	3239	3076	3226
Number of speakers	35	44	65
Number of words	6677	6437	6529
Different words	2891	2503	2908

Table 3: Statistics of annotated speech material.

Show 1 and Show 2 were late night news and Show 3 was the evening news. As evening news have more short reports, also the number of speakers is higher than for the late night news. The number of words and number of different words is similar for all shows. There were 1727 different words common to all three shows.

6. Conclusion and future work

The paper presented an ongoing work on generation of Slovenian Broadcast News speech database. After the speech database will be completed, the speech recognition system will be developed.

Acknowledgment

The authors would like to thank Joao P. Neto and Hugo Meinedo from INESC ID, Portugal for their assistance in the initial phase of the project.

7. References

Barras, C., Lamel, L., Gauvain, J.L., 2001. Automatic Transcription of Compressed Broadcast Audio. *Proc. of ICASSP, 2001*, Salt Lake City.

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

Brousseau, J., Beaumont, J.F., Boulianne, G., Cardinal, P., Chapdelaine, C., Comeau, M., Osterrath, F., Ouellet, P., 2003. Automatic Closed-Caption of Live TV Broadcast News in Frenc, *Proc. Eurospeech 2003*, Geneva, Switzerland.

Cole, R., et.al., 1995. The Challenge of Spoken Language Systems: Research Directions for the Nineties. *IEEE Trans. on Speech and Audio Processing*, vol.3, no. 1.

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

Imai, T., Kobayashi, A., Sato, S., Tanaka, H., and Ando, A., 2000. Progressive 2-pass decoder for real-time broadcast news captioning. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp.1937-1940, Istanbul, Turkey.

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.

http://www.etca.fr/CTA/gip/Projets/Transcriber/-fr/user.html

http://www.ldc.upenn.edu/Projects/Corpus-Cookbook/-transcription/broadcast-speech/english/index.html

Martinčič-Ipšič, S., Žibert, J., Ipšič, I., Mihelič, F., 2002. Speech Recognition of Slovenian and Croatian Weather Forecasts, *Information Society IS'2002, Language Technologies* Ljubljana, Slovenia.

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*, Volume 37, Issues 1-2, 1:3–14.

Sepesy Maučec, M., 2001. *Adaptacija jezikovnega modela na vsebinsko specifično besedišče*, Ph.D. thesis, University of Maribor, Slovenia.

Toporišič, J., et.al., 2001. *Slovenski pravopis*. Ljubljana.

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