Development of Slovenian Broadcast News Speech Database

Janez Žibert, France Mihelič

Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
{janez.zibert, mihelicf}@fe.uni-lj.si

Abstract

The paper reviews the development of a new Slovenian broadcast news speech database. The database consists of audio, video and annotation transcripts of about 34 hours of television daily news program captured from the public TV station RTVSLO. The paper addresses issues concerning transcription and annotation of the collected data, provides information on content analysis and basic statistics of the collected material and reports about preliminary evaluation of automatic segmentation.

1. Introduction

Nowadays there is a significant need to deal with large amounts of multimedia information resulting in the growing demand to shift content-based information retrieval from text to various multimedia sources.

One of the major sources of multimedia information are broadcast news (BN). In the last few years there has been increasing interest in the research of spoken document retrieval from BN including automatic transcription, topic detection and indexing (Makhoul et al., 2000; View4You, 1998; Woodland, 2002; Chen et al., 2002; Gauvain et al., 2002; Beyerlein et al., 2002; Robinson et al., 2002).

The processing of broadcast radio and television news poses a number of challenges for information retrieval systems based on large-vocabulary continuous speech recognition. The data in broadcasts are not homogeneous and include a number of data types for which speech recognition systems trained on read speech corpora have high error rates (Woodland, 2002).

In order to develop technology for the processing of BN data in Slovenian language, we had started to collect a Slovenian BN database (SiBN). The database in current state consists of audio, video and annotation transcripts of about 34 hours of television daily news program from public TV channel RTVSLO-1.

Our goal was to obtain a representative corpus of the Slovenian broadcast news speech data collected under realistic conditions. In contrast to other read speech databases in Slovenian language (Mihelič et al., 2003) this database consists of multiple speakers, variable speaking styles, variable recording conditions in the presence and variation of background noises, covering different type of news stories.

The SiBN database will be used for the development of a BN speech recognition system and a system for topic detection and indexing of the Slovenian broadcast news.

In the next section we are reviewing our work on corpus development: data collection, transcription annotation followed by description of some conversion tools. In section 3 issues concerning database evaluation are addressed and preliminary results of automatic segmentation are provided.

2. Development of the database

2.1. Data Collection

The SiBN database consists of one-hour long daily-news TV shows provided by the national broadcast company RTVSLO. We decided to capture 7PM daily news program Dnevnik, which is one of the most popular daily-news programs in Slovenia according to TV ratings. The program provides different type of international, national and local news and special broadcasts dedicated to sports, financial, cultural news and weather reports. Table 1 summarizes the recorded broadcasts by type and time duration of the news.

Program	Duration	Type
Dnevnik I	10:59	Slovenian news
	3:53	international news
Dnevnik II	4:52	local news
Denar	0:31	finance and business news
Sport	3:18	sport news
Vreme	1:38	weather reports
Magnet	1:43	cultural news
Total	29:14	

Table 1: Time duration in hh:mm format of different types of broadcasts in the SiBN database.

We had collected 34 one-hour long broadcasts from May to August 2003. They were captured using Pinnacle PCTV-Pro TV card. The recordings comprised of audio, video and teletext data. Audio data were sampled at a 16000 Hz sample rate and stored as 16-bit PCM encoded mono waveform audio files. Video files were compressed and stored in Windows Media Video format. Teletext data includes subtitling text captured simultaneously during recording of the broadcast news.

2.1.1. Subtitling via Teletext

The public TV channel SLO that served as the source of data/information provides in addition subtitling information via teletext. The recorded broadcasts were not fully subtitled, only some portion of the daily news included subtitling transcriptions. It was estimated that teletext data provided about 40% of transcriptions on the word level.

The subtitling text stream is reasonably well aligned in time with audio content but due to the differences between speech rate and rate of text display, it was not uncommon to find words and phrases that were spoken but shortened or omitted from captions. The teletext data also did not include markers that flag topic and speaker changes.

We additionally encountered one major problem: teletext decoder simply maps diacritics onto their corresponding non-diacritical letters; i.e. 'č' become 'c', 'š' is replaced with 's' and 'ž' with 'z'. Thus, we had to insert diacritics at right positions. To accomplish this, we designed automatic text conversion tool based on Aspell spell checker[1] for Slovenian language.

Teletext data were further processed into the appropriate format for transcriptions, where we had to manually add accurate time stamps, insert disfluencies and make annotation of speakers, topics and signal conditions.

2.2. Transcription and Annotation

The transcription and annotation work involved five people with different tasks: four transcribers and one final supervisor. Before start annotators were trained on one-hour-long BN recordings. While the transcribers had to provide signal segmentation and correct transcription and annotation, supervisor had to check transcription alignment, annotation structure, speaker and topic consistency and spelling errors.

Annotation process was performed in several passes. In the first pass annotators were instructed to focus simply on establishing time stamps for important and useful points in the recordings, such as story boundaries, speaker turns, and convenient breaks (e.g. breath pauses) for splitting long turns into manageable chunks. Once this was done a separate pass for typing a fully accurate record of what was spoken was performed. In the next pass additional annotation was needed to mark speaker and topic attributes, background and other signal conditions. In this way the annotator can focus more carefully on a smaller set of decisions that needed to be made for a particular stage of annotation and therefore could more efficiently spot and correct errors made in previous passes, as is pointed out in (Graff, 2002).

For one-hour-long broadcast annotators had spent on average 12 to 15 hours for transcription process and additionally 2-3 hours for supervising the transcriptions. The total time spent for transcription of 1h broadcast was greatly reduced when the teletext data was included in the pre-processed transcription.

The transcription process was performed using the Transcriber tool (Barras et al., 2001) following the LDC Hub4 broadcast speech transcription conventions[2]. The data format of transcriptions follows the XML standard with Unicode support (for Slovenian language).

2.3. Conversion and Verification Tools

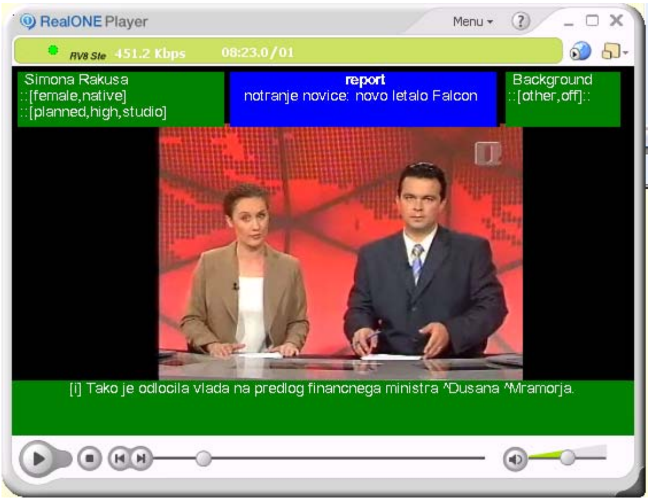
Transcriber transcription format (TRS) is similar to universal transcription format (UTF) proposed in (NIST, 1998) with minor differences in treating speaker characteristics and naming entities which are only optionally present in UTF (Barras et al., 2001).

However, in order to follow Hub-4 evaluation specifications for speech recognition accuracy we had to provide conversion from TRS/UTF speaker, speech and channel attributes into the focus conditions (Pallett, 2002).

The conversion algorithm performs modification of the speaker dialect (native or nonnative), speaking mode (planned or spontaneous), channel bandwidth (telephone or studio), fidelity (high, medium or low), and different background conditions (music, speech, shh, other) as provided by the TRS format to 7 focus conditions defined in (Pallett, 2002). The conversion from speaking mode, channel bandwidth and fidelity to corresponding focus conditions is straight forward, while a special care is needed for modification of speaker dialect, different type of background conditions and in the presence of overlapping speech.

Based on the conversion algorithm we developed a tool for parsing Transcriber XML data and converting it to NIST Sclite segment time mark format (STM)[3]. The conversion tool is slightly different than those bundled in the Transcriber toolkit.

The STM file identifies time intervals along with the information about speaker, focus conditions and transcription for those intervals. The STM format will be used as the 'base' format in the reference transcriptions in the development of the BN speech recognition system. This format is agreed upon and is to be used in the transcriptions of multilingual BN speech database within the COST278 action (Vandecatseye et al., 2004) in which our research group is an active partner.



Simona Rakusa
::[female,native]
::[planned,high,studio]

report
notranje novice: novo letalo Falcon

Background
::[other,off]::

[i] Tako je odlocila vlada na predlog financnega ministra ^Dusana ^Mramorja.

Figure 1: Audio-visual representation of annotated data using SMIL-enabled video player.

We had additionally developed a tool for an audiovisual inspection of the annotated data. This tool integrates audio and video streams with transcription texts using synchronized multimedia integration language (SMIL)[4]. As it

[1] GNU Aspell: http://aspell.sourceforge.net/

[2] LDC transcription conventions for Hub-4 English: http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/conventions.html

[3] NIST Sclite input file format: ftp://jaguar.ncsl.nist.gov/current_docs/sctk/doc/infmts.htm

[4] The Synchronized Multimedia Integration Language: http://www.w3.org/AudioVideo/

can be seen form Figure 1 a supervisor can check transcription alignment, speaker attributes, background conditions and topic descriptions along with audio-video stream using SMIL-enabled audio/video player.

3. Evaluation of the database

Evaluation of the SiBN database was focused on the two issues: content richness and consistency of annotations. in addition, automatic segmentation was performed in order to check manual signal segmentation and transcription alignments.

3.1. Content Analysis

A desirable feature of the database is to be rich in terms of acoustic and linguistic content. Hence, statistics related with these concepts have been extracted and analyzed.

Total time duration of the collected data is 34h12m from which 27h53m corresponds to reports and 2h to fillers (8% of the transcribed data). The rest 4h18m of data belongs to jingles, commercial breaks and foreign-language speech, which was not transcribed.

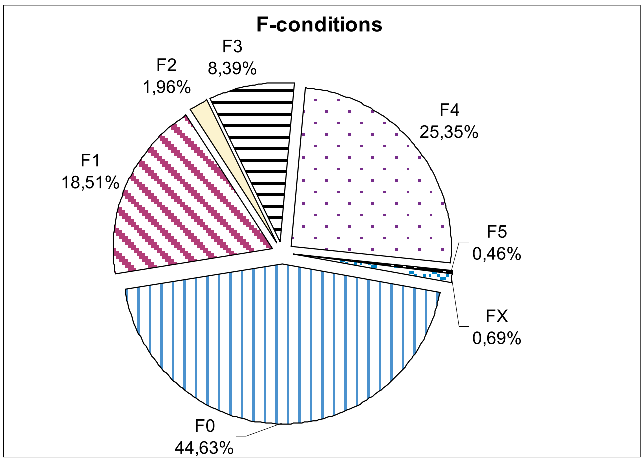


Figure 2: Distribution of speech according to the focus conditions.

The acoustic variability was measured with a set of focus conditions (Pallett, 2002) extracted from SiBN corpus using conversion algorithm described in the previous section. Statistics of focus conditions revealed expected proportion (44%) of the baseline (F0) and spontaneous (F1) speech (19%), Figure 2. Approximately half an hour of speech originating over telephone channels (F2) represents 2% of the speech data, which is relatively a small proportion of material compared to similar BN databases (Federico et al., 2000; Pallett, 2002). On the other hand, the SiBN database contains a considerable amount (8%) of speech with the presence of background music (F3) due to the fact that almost all headline news in most of the fillers and broadcast shows with cultural news have music in background. Statistics also yielded a substantially great proportion of material (25%) in the degraded acoustic conditions (F4) and lower proportion of material in the miscellaneous conditions (FX) (less than 1%). The speech database includes a relatively small amount of speech from nonnative speakers (F5) (less than 1%).

Another considerable issue concerning acoustic variability is the number and distribution of speakers represented in the database. The total number of speakers is 1477. The majority of speakers (1166) belongs to the native speakers' group and the rest are nonnative or foreign-language speakers. The database includes 1113 male and 346 female speakers who produced 41% of the speech material calculated in time duration of utterances.

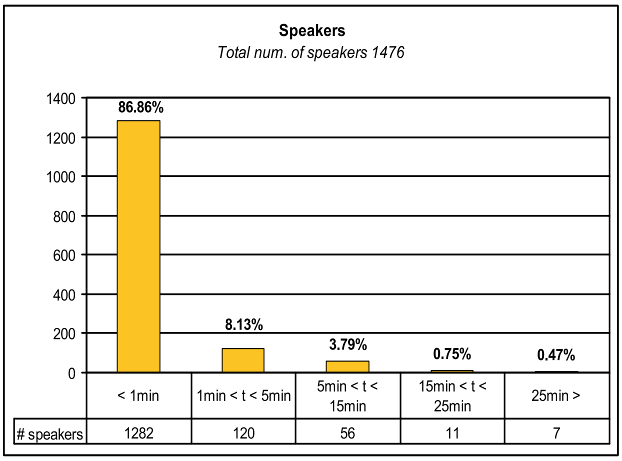


Figure 3: Distribution of speakers according to the available speech in time.

Figure 3 presents the distribution of speakers according to their available speech in time. Just according to our expectations, the majority of speakers belongs to the first group in which given amount of speech is at least available. On the other hand, there are 6 anchor speakers each of whom providing approximately 1h of speech.

The linguistic content was analyzed addressing the topics and word level. The SiBN corpus consists of 877 different topics and 148 filler sections. Reports covering local news represents 57% of data, international news 16%, sport news 12%, finance and business news 2%, cultural news topics 7% and weather reports 6% of all data in the corpus. This statistics is based on a time duration of each section.

The transcripts contain 255K words for a vocabulary size of 38K words. Additional statistics was made for non-speech events, where the major contribution of various types of breath noises was noticed.

3.2. Automatic Segmentation

Automatic segmentation experiments were performed in order to check manual signal segmentation and transcription alignments. The goal of automatic segmentation of audio signal is to detect changes in speaker identity, environmental condition, and channel. The problem was to find acoustic change detection points in an audio stream.

Segmentation was performed using the DISTBIC method (Delacourt et al., 2000), which is a two-pass change detection technique. In the first pass we used generalized likelihood ratio (GLR) to determine the turn candidates. Gaussian probability density function parameters were estimated for a two-second-long adjacent windows placed at every point in the audio stream. When the GLR distance between bordering windows reached a local maximum, a segment boundary was generated. In the second pass Bayesian

Information Criterion (BIC) was applied to validate or discard candidates from the first pass, (Chen et al., 2002).

The result of a segmentation could assume two likely types of error that could be measured by precision (PRC) and recall (RCL). They are defined as (Kemp et al., 2000):

$$RCL = \frac{\text{number of correctly found boundaries}}{\text{total number of boundaries}}$$

$$PRC = \frac{\text{number of correctly found boundaries}}{\text{number of hypothesized boundaries}}$$

A hypothesized segment boundary t is judged as correct, if it lies within the time interval $t_0 - \Delta t < t < t_0 + \Delta t$ of the reference boundary t_0.

In our experiments we chose $\Delta t = 1.0$ s. All other parameters of the segmentation system were tuned on one-hour-long broadcast. Evaluation results of automatic segmentation are shown in Table 2

TNB	NHB	NCB	INS	DEL	RCL	PRC
6370	5926	4621	1305	1749	0.73	0.78

Table 2: Automatic segmentation results on the SiBN data, where TNB means total number of boundaries to detect, NHB number of hypothesized boundaries, NCB number of correctly found boundaries, INS number of inserted boundaries and DEL number of deleted boundaries. RCL and PRC are recall and precision respectively.

4. Conclusion

The issues concerning transcription and annotation of the collected data were addressed in the paper, where we followed Hub-4 annotation instructions and LDC transcription conventions. We also produced some tools for conversions between different transcription formats and for audio-visual inspection of the annotated data. Content analysis and basic statistics were performed on the collected material in order to explore the acoustic and linguistic variability encompassed in the database. Additionally, preliminary evaluations and automatic segmentation of the data was made to check and improve the transcription alignments and consistency of the annotations. The achieved results document a good feasibility and speak in favor of the future applications.

The SiBN database will be used for the development of a speech recognition system for automatic transcription and a system for topic detection and indexing of the Slovenian broadcast news.

5. Acknowledgment

The authors would like to thank the public broadcasting company RTVSLO for giving us permission to freely use the recorded material for scientific purposes.

6. References

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

Beyerlein, P., Aubert, X., Haeb-Umbach, R., Harris, M., Klakow, D., Wendemuth, A., Molau, S., Ney, H., Pitz, M., Sixtus, A., 2002. Large vocabulary continuous speech recognition of Broadcast News - The Philips/RWTH approach. *Speech Communication*, Volume 37, May 2002, 1:109–131.

Chen, S. S., Eide, E., Gales, M. J. F., Gopinath, R. A., Kanvesky, D., Olsen, P., 2002. Automatic transcription of Broadcast News. *Speech Communication*, Volume 37, May 2002, 1:69–87.

Delacourt, P., Wellekens, C. J., 2000. DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Comm.*, Volume 32, September 2000, 1:111–126.

Federico, M., Giordani, D., Coletti, P., 2000. Development and Evaluation of an Italian Broadcast News Corpus. *Proc. of the LREC 2000*, Athens, Greece, Volume 2, 2:921–924.

Graff, D., 2002. An overview of Broadcast News corpora. *Speech Comm.*, Volume 37, May 2002, 1:15–26.

Gauvain, J. L., Lamel, L., Adda, G., 2002. The LIMSI Broadcast News transcription system. *Speech Communication*, Volume 37, May 2002, 1:89–108.

Kemp, T., Geutner, P., Schhmidt, M., Tomaz, B., Weber, M., Westphal, M., Waibel, A., 1998. The Interactive Systems Labs View4You video indexing system. *Proceedings ICSLP 98*, Sydney, Australia, December 1998, Volume 4, 4:1639–1642.

Kemp, T., Schmidt, M., Westphal, M., Waibel, A., 2000. Strategies for Automatic Segmentation of Audio Data. *Proc. of the ICASSP 2000*, Volume 3, 3:1423–1426.

Mihelič, F., Gros, J., Dobrišek, S., Žibert, J., Pavešič, N., 2003. Spoken language resources at LUKS of the University of Ljubljana. *International Journal of Speech Technology*, Volume 6, 3:221–232.

Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., Srivastava, A., 2000. Speech and Language Technologies for Audio Indexing and Retrieval. *Proc. of the IEEE 88*, Volume 8, 2000, 8:1338–1353.

NIST, 1998. Universal transcription format (UTF) annotation specification for Evaluation of Spoken Language Technology Corpora. *http://www.nist.gov/speech/tests/bnr/bnews_99/utf-1.0-v2.ps*

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, May 2002, 1:3–14.

Robinson, A. J., Cook, G. D., Ellis, D. P. W., Fosler-Lussier, E., Renals, S. J., Williams, D. A. G., 2002. Connectionist speech recognition of Broadcast News. *Speech Comm.*, Volume 37, May 2002, 1:27–45.

Vandecatseye, A., Martens J. P., Neto, J., Meinedo, H., Garcia-Mateo, C., Dieguez. J., Mihelic, F., Zibert, J., Nouza, J., David, P., Pleva, M., Cizmar, A., Papageorgiou, H., Alexandris, C., 2004. The COST278 pan-European Broadcast News Database. *Proc. of the LREC 2004*.

Woodland, P. C., 2002. The development of the HTK Broadcast News transcription system: An overview. *Speech Comm.*, Volume 37, May 2002, 1:47–67.