Slovak Broadcast News Speech Corpus for Automatic Speech Recognition

Matúš Pleva, Jozef JUHÁR, Anton ČIŽMÁR

Department of Electronics and Multimedia Communications,
Technical University of Košice, Park Komenského 13, 041 20 Košice,
Slovak Republic
Email: Matus.Pleva@tuke.sk, Jozef.Juhar@tuke.sk, Anton.Cizmar@tuke.sk

Abstract

One of the most attractive applications of automatic speech recognition is automatic transcription of broadcast news archives. The complete automatic transcription system is able to search the needed information in metadata database of TV or radio shows.

This paper deals with the algorithm of creating the broadcast news speech corpus, which could be used for training the acoustic and language models for automatic speech recognition engines in Slovak language. It contains the statistics of the new KEMT-BN broadcast news database.

The description of the standard processes of recording, transcribing, converting and storing all data and using them for testing and training the broadcast news transcription system is here presented. Perl scripts are used for making the whole process as automatic as it is possible.

1. Introduction

After successful contribution to the COST278-BN database [1] an initiative to continue in recording and transcribing broadcast news materials in Slovak language started. The resulting database will be used for development and evaluation of the automatic broadcast news transcription system, and all of his modules.

2. Collecting the database

The database was recorded from Slovak Television channel - STV 1 using Technisat Airstar PCI card from DVB-T broadcast on channel 44 in Košice region. The MPEG2 Transport stream is then truncated using Mpeg2Cut freeware, which removes the recording parts before and after the news broadcast. Then the audio channel is de-multiplexed from the stream using MPEG Tools from TMPGEnc software resulting .mp2 file (48kHz stereo 224kbps CBR).

Next the audio file needs to be decompressed to wav file using for example Winamp or freeware foobar2000 tool. The resulting wav file is then converted to mono and down-sampled to 16kHz using SoX freeware tools (sox <infile> -r 16000 <outfile> polyphase).

The complete video recording is then converted also to real media format with a resolution of 352x288 because of reducing the file size. The video recording is important when transcribing the speaker names and topics descriptions.

3. Transcription process

Transcription process consists of manual orthographic transcribing of the whole audio

recording using generated wav file and Transcriber freeware tool. The annotation process follows the LDC transcription conventions for HUB4 [2]. Resulting file is xml based .trs file.

After completing the transcription the .stm file is generated. The .stm file is the source format for next processing of the recordings, as segmentation, and conversion to other speech database standards [1].

3.1. Special transcription rules

The major **speaker turn attributes** were channel (studio/telephone) and fidelity. Fidelity low/medium/high has different meanings for different channel conditions [3].

For the studio speech, *high* fidelity is used for conversations that take place inside a studio. Usually, this is when the anchor person is talking or when a video story is commented by a journalist that is recorded in a studio.

Medium fidelity refers to speech that is captured in the field, usually situations where the journalist is making a street interview.

Low fidelity refers to situations where there is noise in the transmission channel. In the case of telephone speech, *high* refers to clear (clean) speech, *medium* to noisy speech that is still easy to understand though, and *low* to speech that is difficult to understand [4].

Table 1 summarizes this coding scheme.

"Table 1. "Coding of channel and fidelity attributes"

		Channel	
		Studio [Bandwidth > 4kHz]	**Telephone [4kHz Bandwidth]**
Fidelity	**Low**	Channel noise	Not intelligible
	Medium	Field	Noisy
	High	Studio	Sounds clear

The **speech utterances** should not be too long and every speaker inspiration event should be regarded as a potential breakpoint.

When a **silence inside a speaker turn** is less than 0.5 seconds it is not marked at all. When it is between 0.5 and 1.5 seconds, a breakpoint in the middle of the silence is inserted. When the silence is longer than 1.5 seconds, two breakpoints delimiting the silence are inserted.

The **sections blocks** are categorized as reports (news stories), fillers (headlines and short story descriptions) and nontrans (commercials and jingle segments).

All **jingle segments** are identified as such and marked by a noise event tag. When the TV station uses different jingles at the beginning and the end of a show, each jingle gets an additional suffix indicating its begin/end category.

Foreign language utterances are marked with language event tags and are not transcribed.

Interjections and non-lexemes was defined so as to pursue that the same tags are used for the same words/sounds in all the data.

3.1. Using Transcriber

Transcriber allows to mark speech and noise time regions, mark a report or filler, mark the speaker identification (known speakers from the video recording titles or unknown speaker as speaker#1, etc.), noise events and transcribed text (see Figure 1).

Transcriber allows also mark a foreign speech, coding of channel and fidelity of speaker turn (see Table 1.).

Transcriber has default western encoding and French noise labels, so is important for transcribers to load a Slovak configuration file. This is problem because the configuration file has a strict format, and gives many errors when loading a edited one, which can not be equal for all users.

Transcriber allows export of the transcription XML .trs format to .stm, LDC .typ, text, LIMSI .lbl and HTML format.

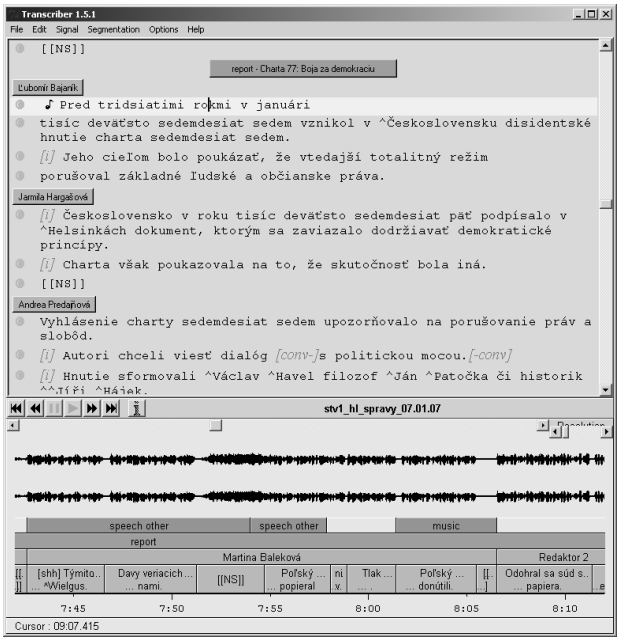

Figure 1. Transcriber window

Transcriber is available for Windows and also Linux and MacOS X. Transcriber is developed with the scripting language Tcl/Tk and C extensions.

Transcriber supports spell checking, but is not functional for Slovak language due to encoding problems. Spell checking was done in MS Word and then all word errors was corrected again in Transcriber. Solve this problem will be a big advantage in the future work.

4. Statistics across the KEMT-BN database

The collected transcribed broadcast news material consists of approximately 52 hours of different recordings collected during 4 years. There was 5 transcribers working on the database and the amount of transcribed data from each of them is depicted on the Table 2 below.

"Table 2. Collected broadcast news material"

Transcriber	Duration [h:m:s]
Pleva (TA3) [5]	03:10:37.021
Kovacik (STV1) [6]	10:26:00.29
Svirloch (STV1) [7]	05:42:11.247
Solc (STV1) [8]	15:05:01.04
Malik (STV1) [9]	17:18:16.469
Summary	51:42:06.067

The statistics of the resulted dictionary is:

- Total 371636 words
- 47050 words in dictionary
- 8239 words outside dictionary (not complete or foreign language words or indexes)

In order to measure and compare speech recognition accuracy for some of the acoustical conditions, the database was segmented into the following set of focus conditions [10]:

- F0: baseline broadcast speech, the baseline condition, includes prepared speech recorded in studio conditions
- F1: spontaneous broadcast speech includes spontaneous speech recorded in studio conditions
- F2: speech over telephone channels includes speech collected under reduced-bandwidth conditions
- F3: speech in the presence of background music includes prepared and spontaneous speech at an SNR of 10 to 20 dB, A-weighted
- F4: speech under degraded acoustical conditions includes prepared and spontaneous speech degraded by additive noise, environmental noise, or nonlinear distortions, at an SNR of 10 to 20 dB, A-weighted
- F5: speech from non-native speakers includes studio-quality intelligible English speech spoken by non-native speakers of American English

(including English spoken by natives of the United Kingdom)

- FX: miscellaneous includes speech that does not satisfy any of the above conditions, or speech that simultaneously satisfies more than one of the conditions F1 through F5 (such as non-native speech with music in the background)

Statistic of the focus conditions measured in the database is in the Table 3 below:

"Table 3. Collected broadcast news material according to focus conditions"

Focus condition	Duration [h:m:s]
F0	20:55:16.892
F1	09:34:18.995
F2	00:52:28.889
F3	05:21:16.377
F4	08:20:55.086
F5	00:53:12.024
FX	02:30:33.545
No speech	03:07:09.792

7. Conclusion

The new KEMT-BN database contain approximately 52 hours of broadcast news transcribed recordings with approximately 21 hours of clean planned studio speech. This feature makes this database a very useful corpus, not only for training but also for testing and evaluating purposes of many kinds of applications

This database is now used for training and testing the speech detection, speaker detection, speaker clustering, gender detection and continuous speech recognition engines developed in our department.

8. Appendix and acknowledgments

The work presented in this paper was supported by the Ministry of Education of Slovak Republic under research projects AV 4/0006/07, VEGA No. 1/4054/07 and by the Grant agency APVV project No. APVT-20-029004.

References

[1] A. Zgank, Z. Kacic, A. Moreno, M. Caballero, F. Diehl, K. Vicsi, G. Szaszak, J. Juhár, S. Lihan: "The COST 278 initiative – crosslingual speech recognition with large telephone database", *Proc. LREC' 2004 – 4th International Conference On Language Resources And Evaluation*, Lisabon, 26-28 May 2004, pp. 2107–2110.

[2] http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/index.html

[3] A. Vandecatseye, J. P. Martens, J. Neto, H. Meinedo, C. G. Mateo, J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A.Cizmar, H. Papageorgiou, Ch. Alexandris: "The COST278 pan-European Broadcast News Database", *Proc. LREC' 2004 – 4th International Conference On Language Resources And Evaluation,* Lisabon, 26-28 May 2004, pp. 873 - 876.

[4] J. Zibert, F. Mihelic, J. P. Martens, J. Neto, H. Meinedo, C. G. Mateo, L. Docio, J. Zdansky, P, David, M. Pleva, A. Čižmár, A. Zgank , Z. Kacic, C. Teleki, K. Vicsi: "The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results"*, Proc. Interspeech 2005 (Eurospeech) – 9th European Conference on Speech Communication and Technology,* Lisabon, September 4-8 2005, pp. 629 – 632.

[5] M. Pleva, J. JUHÁR, A. Cizmar: "About development and evaluation of multilingual database for automatic broadcast news transcription", *Acta Electrotechnica et Informatica*, Vol. 4, No. 2, 2004, pp 56 – 59.

[6] L. Kovacik: "Corpus of Broadcast News Recordings", Master Thesis, KEMT FEI TU Kosice, 2005.

[7] M. Svirloch: "Corpus of Broadcast News Recordings", Master Thesis, KEMT FEI TU Kosice, 2006.

[8] M. Solc: "Creating and testing corpus of TV/R news speech records for the needs of automatic detection of speech", Master Thesis, KEMT FEI TU Kosice, 2007.

[9] P. Malik: "Corpus of Broadcast News Recordings", Master Thesis, KEMT FEI TU Kosice, 2007.

[10] T. Hain, P. C. Woodland: "Segmentation And Classification Of Broadcast News Audio." *Proc. ICSLP'1998 - 4th International Conference on Spoken Language Processing*