Matus Pleva – Jozef Juhar *

BUILDING OF BROADCAST NEWS DATABASE FOR EVALUATION OF THE AUTOMATED SUBTITLING SERVICE

This paper describes the process of recording, annotation, correction and evaluation of the new Broadcast News (BN) speech database named KEMT-BN2, as an extension for our older KEMT-BN1 and COST-278 databases used for automatic Slovak continuous speech recognition development. The database utilisation and statistics are presented. This database was prepared for evaluation of the automated BN transcription system, developed in our laboratory, which is mainly used for subtitle generation for recorded BN shows. The speech database is the key part of the acoustic models training for specific domains and also for speaker and anchor adapted models creation.

Keywords: *Broadcast news, segmentation, speech recognition, Transcriber.*

1. Introduction

The development of continuous speech recognition (CSR) systems in Slovak language expects a large amount of different language resources to be collected [1 and 2]. First of all, the speech database needs to be built, which is also the most expensive and demanding task [3]. The building of the textual database of Slovak texts for language modelling is also challenging [4 and 5] and could be done using modern crawling technologies and post-processing, morphological analysis etc. [6].

The KEMT-BN2 database campaign was carried out between 2009 and 2011. It consists of broadcast news (BN) shows from the first Slovak public service broadcaster television (STV1 – Jednotka). The transcription task was mainly realized by brigadiers, and then trained and evaluated by transcription specialist. The database is a follower of the KEMT-BN1 [7] and the Slovak part of the COST-278 [8] database realized in our laboratory [9] (recorder from TA3 news television).

The purpose of the specialized BN databases is to build and evaluate the automatic transcription system for BN shows [10]. This system should have special BN acoustic models for different types of speech in BN shows (F-conditions) [11], special acoustic models for anchors or speakers with high occurrence in the news (politicians, sportsmen, artists, etc.) [12, 13 and 14] and also a special language model from the BN domain [15 and 16]. To use these models in the special detection system for speakers, different types of speech etc. should be provided [17 and 18].

This paper describes the process of recording and collecting the audio materials of the BN shows. Next, the transcription process and the evaluation of the transcriptions are presented. Finally the statistics of the collected annotated data in the database is depicted and discussed in conclusions and future work.

2. Recording the shows

The broadcast news shows were recorded from DVB-T channel multiplex streaming the PS (program stream) data to the disk using Technisat Airstar PCI card [19] from testing broadcast on channel 25 in Kosice region before an official digitisation process. The MPEG2 Program Stream was captured with time reserve, but it was not truncated because a jingle detection algorithm based on Euclidian distance or DTW is planned to be developed and evaluated on this data later.

The audio subchannel is de-multiplexed from the stream using DGMPGDec DGIndex [20] GPL (GNU Public License) licensed software resulting in *.mp2* file (*48 kHz* stereo *128 kbits* CBR constant bitrate quality MPEG-1 Audio Layer 2 codec).

Next, the audio file needs to be converted to a format compatible with transcription software and delivered to annotators (the file size is also important). The used *Transcriber* software [21] mentioned in the next chapter has several bugs when using mp3 format (the time was not correlating with wav or video) so the *mp2* files (not compliant) were recompressed to *ogg* format (Ogg Vorbis *160 kbps* q5.0 mono) using the freeware foobar2000 tool [22] with *Vorbis* plugin. After that also a mono PCM *16 kHz wav* file was decompressed for database utilisation purposes.

The complete video recordings is planned to be converted also to a well supported video streaming format for using in web application for presenting the database with captions for the public

* **Matus Pleva, Jozef Juhar**
Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Slovakia, E-mail: Matus.Pleva@tuke.sk

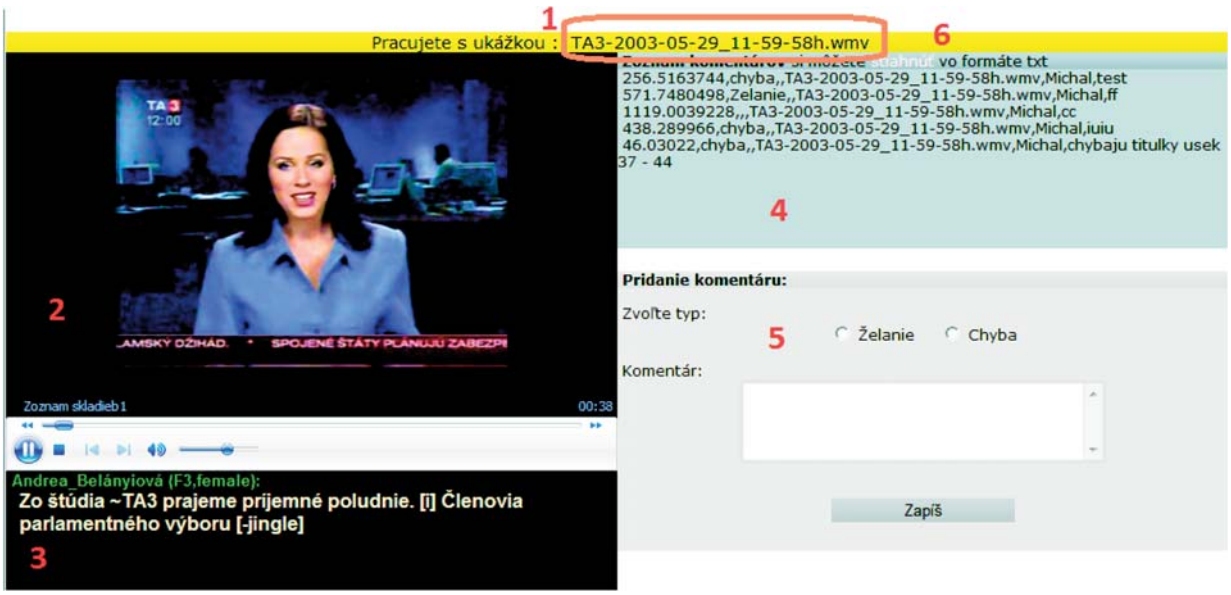

Fig. 1 Broadcast news shows (1/2) transcriptions (3) presentation on web interface (6) with ability to send an error report (5) to the administrator with automatic timestamp (4) of the paused video

(Fig. 1 – the COST278 TA3 part of the database on the web) [23]. The video recording is important when transcribing the speaker names (from captions in the video) and topics descriptions too.

3. Transcription of the speech and non-speech audio events

Transcription process consists of manual orthographic transcriptions of the whole audio recording using *Transcriber 1.5.1* tool – a free software under GPL license (Fig. 2) [21]. The annotation process follows the LDC (Linguistic Data Consortium) transcription conventions for HUB4 [24] (DARPA-sponsored Hub4 continuous speech recognition evaluation) extended using new rules for Slovak language and future use for lexical and language modelling. The native *xml* file format file is *.trs* file.

A) STM export

After completing the transcriptions the .stm (the NIST Scoring toolkit Sclite [25] – a more simple text file format exported from Transcriber) file is generated. The .stm file is the source format for next processing of the recordings, as segmentation and conversion to other speech database and online subtitles standards [26] which are suitable for using reference speech recognition training procedure described in [27]. We developed a special set of Perl scripts for conversion from *wav* and *stm* file pairs to the standardized SpeechDat database format for this purpose [28].

B) Transcriber modifications

The Transcriber toolkit was slightly modified for these purposes. The description of noise markers was translated and extended (the annotators have to enter the noise marker/tags only using menu – to avoid frequent typos in non-speech tags).

Next, also the conversion script for *stm* format export was modified to include all tags in resulted *stm* file (some of them were filtered).

Finally, the Slovak spellchecking feature was realized using free GPL licensed Aspell (http://aspell.net/) dictionary and modifying the corresponding spelling TCL/Tk script (which should send only the words to the dictionary – not tags) which was not

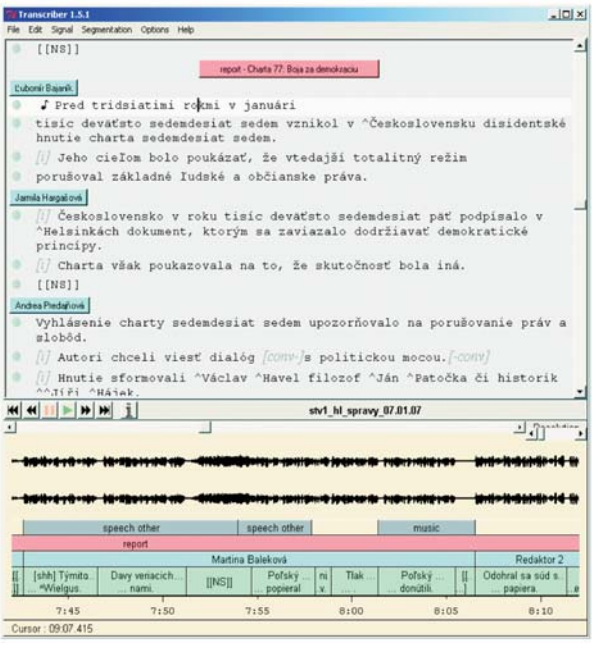

Fig. 2 Transcriber window with audio segments, visualized waveforms, transcribed text, speaker and topic tags

able to work with Slovak symbols (CP1250 or UTF-8 encoded characters) in the text.

C) Segmentation and foreign languages

The speech utterances in the database have not been too long and every speaker inspiration event (breathing – tint) should be regarded as a potential breakpoint.

According to segmentation, the silence inside a speaker turns shorter than 0.5 seconds was not marked at all. Breakpoint in the middle was inserted when the pause in the speech utterance is between 0.5 and 1.5 seconds. When the pause was longer than 1.5 seconds, a special silence segment was inserted [7].

Foreign language utterances were marked with language event tags and should not be transcribed at all.

4. Database corrections

The database includes many typos, mistakes, misspelling and strange characters also after second annotator review of every transcription. The correction process is important because every wrong annotation could decrease the quality of the resulting acoustic or language models.

The process used in our laboratory for acoustic models training is very sensitive to every discontinuity in the database (refrec – Reference Recognizer from COST249) [27]. This process is also affected using the conversion scripts from file pairs (*wav* + *stm*) to SpeechDat format, including the mapping of the noise markers/tags and generating the phonetic lexicon.

There are more crucial points in the acoustic models (AM) training procedure described in [29]:

- Generating the *word level phonetic transcription* of all segments which will be used in the training procedure. Usually the script finished with error that *some word was not found in the phonetic lexicon*, generated during the conversion of the database.
  - Mainly it is a problem that some tags or non-word units passed to the training because the *tags mapping* (from huge set of noise tags to simple SpeechDat [sil] [spk] [sta] [int] tags [26]) *missed some new tag* (the developer needs to decide how to map it). Also typing errors (typos) are discovered during this stage like: missing character or mistyped character.

- Generating *initial monophone models*. Sometimes there is a problem that *for specific segment a proper label file is missing*.
  - This error is caused by *inconsistency of the two* filtering and index file generating *scripts*, when one script filters out a segment as not suitable for training (and do not include its labels in the master label file) but the script for generating the file-list of training segments decide that this segment could be used for training. The architecture of the used training procedure should be changed in the future to use only one set of filtering rules in both scripts and including them from specialized configuration file.

- Generating the *phone prototypes*. During this stage the developer sometimes discovers that an *unknown phoneme (or unwanted) is in the training* or some phoneme or noise *model* is missing.
  - This error is usually caused by a *non-Slovak word* (should not be annotated or should be marked with a special tag) or *filtering script error*, when some tag was filtered (during training set generation) like not important for AM training (lexical tags), but then we found out that also another important noise tag was filtered.

- Generating *tied triphone models* sometimes crash because some *phoneme is missing in the decision tree* or phonemes class definitions.
  - Sometimes we want to try a new phoneme set (reduced or more specific) for testing the impact of the precise phoneme definitions on the resulting system. During this stage sometimes the phoneme class definition should be changed or the *phoneme mapping is not properly defined* and should be corrected.

- Automatic *forced alignment* errors. When the *forced alignment procedure could not find a suitable automatic alignment* for the segment and its corresponding annotation the segment will be included to the outliers list and will be discarded from the training procedure.
  - Checking this outliers list and *reviewing the original file* and the corresponding *annotation*, the annotation should be corrected because there is usually some error in annotation (sometimes missing word or another word with similar meaning instead the right one – it is complicated to find this type of errors for the annotators because the brain is doing some automatic correction sometimes during monotonous work).

5. Database statistics

The database consists of 291 TV shows in 210 hours of material (including time reserve before and after). The total transcribed database includes *141 hours of annotated audio material* (1'169'832 words in 131'884 speech utterances) and the distribution of Focus conditions is depicted in Table 1 below.

The dictionary generated from this database consists of 95'376 Slovak words and 19'425 foreign words/names, noises, not correctly spelled words, partial/misspelled words or abbreviations.

The phonetic transcription (pronunciation lexicon) was generated using our developed Perl tool, and it is a very important element of the database. The phoneme description is based on SAMPA format [30] standard. Recently we found out that the phonetic transcription based on words for SpeechDat databases is not suitable for sentences, so we decide to change the training script to accept also whole sentences phonetic transcription for better inter-word phonetics.

Speech utterances distribution in the KEMT-BN2 database — Table 1

Focus conditions of the utterances	
F0 – prepared speech in studio	73.46 h
F1 – spontaneous speech in studio	23.13 h
F2 – prepared telephone speech (reduced-bandwidth)	1.20 h
F3 – speech with music in background (SNR<10dB)	13.80 h
F4 – speech under degraded acoustical conditions	35.03 h
F5 – speech performed by non-native speaker	0.36 h
FX – combined conditions of types mentioned above	18.89 h

6. Conclusions

The collection of speech databases is the crucial problem when developing an automatic speech recognition engines for different domains and conditions. The broadcast news task is a very popular issue nowadays, because the government regulation specifies the minimal amount of shows with hidden subtitles for hearing impaired spectators.

The new KEMT-BN2 database brings a very important contribution to broadcast news processing. Not only for speech recognition but also for jingle detection, speech detection, speaker/anchor detection (anchor – hosting character in broadcast programs), segmentation, speaker clustering and different specialized noise modelling for domain specific tasks.

The KEMT-BN2 database has 3 times more data in every important parameter than the previous KEMT-BN1 and the Slovak part of COST-278 database together [7].

In the next period, we plan to use this database for building new acoustic models for broadcast news automatic continuous speech recognition, evaluate these models with previous versions (built on KEMT-BN1 and other databases) on new BN domain specific test set. Together with our colleagues we already prepared the new language model (LM) for BN task (adapted from huge universal LM used in our previous projects [4]). We plan also to implement the sentence level phonetic transcription process in the training script.

Acknowledgement

The research presented in this paper was supported by the Research & Development Operational Program funded by the ERDF (ITMS 26220220155) 50% and by 7th Framework Programme EU ICT project INDECT (FP7 - 218086) 50%.

References

[1] JUHAR, J., CIZMAR, A., RUSKO, M., TRNKA, M., ROZINAJ, G., JARINA, R.: Voice Operated Information System in Slovak, *Computing and Informatics,* vol. 26 (6), pp. 577–603, 2007.

[2] NOUZA, J., SILOVSKY, J., ZDANSKY, J., CERVA, P., KROUL, M., CHALOUPKA, J.: Czech-to-Slovak Adapted Broadcast News Transcription System, *Proc. of INTERSPEECH 2008*, pp. 2683–2686, 2008.

[3] PLEVA, M., JUHAR, J., CIZMAR, A.: About Development and Evaluation of Multilingual Database for Automatic Broadcast News Transcription Systems, *Acta Electrotechnica et Informatica (AeI),* vol. 4 (2), pp. 56–59, 2004.

[4] STAS, J., HLADEK, D., PLEVA, M., JUHAR, J.: Slovak Language Model from Internet Text Data, *Lecture Notes in Computer Science*, Vol. 6456 LNCS, pp. 340–346, 2011.

[5] PROCHAZKA, V., POLLAK, P., ZDANSKY, J., NOUZA, J.: Performance of Czech Speech Recognition with Language Models Created from Public Resources, *Radioengineering,* Vol. 20 (4), pp. 1002–1008, 2011.

[6] HLADEK, D., STAS, J.: Text mining and processing for corpora creation in Slovak language, *Journal of Computer Science and Control Systems,* vol. 3 (1), pp. 65–68, 2010.

[7] PLEVA, M., JUHAR, J., CIZMAR, A.: Slovak Broadcast News Speech Corpus for Automatic Speech Recognition, *Proceedings of RTT 2007 conference,* Zilina, p. 4, 2007.

[8] VANDECATSEYE, A. et al.: The COST278 pan-European Broadcast News Database, *Proc. of LREC 2004*, vol. 6, May 2004, Lisbon, pp. 873–876, 2004.

[9] PLEVA, M.: Building European Broadcast News Database, *Proc. of 4. Doktorandska konferencia a SVOS TU v Kosiciach - SCYR 2004*, Kosice, pp. 85–86, 2004.

[10] PLEVA, M., CIZMAR, A., JUHAR, J., ONDAS, S., MIRILOVIC, M.: Towards Slovak Broadcast News Automatic Recording and Transcribing Service, *Lecture Notes in Computer Science: Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction,* vol. 5042 LNCS, p. 158–168, 2008.

[11] JARINA, R., KUBA, M.: Speech Recognition Using Hidden Markov Model with Low Redundancy in the Observation Space, *Komunikacie (Communications)*, Vol. 6 (4), pp. 17–21, 2004.

[12] NOUZA, J. et al.: Making Czech Historical Radio Archive Accessible and Searchable for Wide Public, *Journal of Multimedia*, Vol. 7 (2), pp. 159–169, 2012.

[13] HRIC, M., CHMULIK, M., JARINA, R.: Comparison of Selected Classification Methods in Automatic Speaker Identification, *Komunikacie (Communications)*, Vol. 13 (4), pp. 20–24, 2011.

[14] CERVA, P., NOUZA, J., SILOVSKY, J.: Study on Cross-lingual Adaptation of a Czech LVCSR System Towards Slovak, *Lecture Notes in Computer Science.* Vol. 6800 LNCS, pp. 81–87, 2011.
[15] NOUZA, J., SILOVSKY, J.: Adapting Lexical and Language Models for Transcription of Highly Spontaneous Spoken Czech, *Lecture Notes in Computer Science,* vol. 6231 LNAI, pp. 377–384, 2010.
[16] JUHAR, J., STAS, J., HLADEK, D.: Recent Progress in Development of Language Model for Slovak Large Vocabulary Continuous Speech Recognition, *New Technologies: Trends, Innovations and Research,* Rijeka: InTech, pp. 261–276, 2012.
[17] PLEVA, M., JUHAR, J., CIZMAR, A.: Speech Detection in the Broadcast News Processing, *Proc. of DSP-MCOM 2005.* Kosice, pp. 84–85, 2005.
[18] VAVREK, J.: Audio Content Classification using SVM Binary Decision Trees, *Proc. of SCYR 2012: 12th Scientific Conference of Young Researchers,* May, Herlany, pp. 80–83, 2012.
[19] http://www.technisat.com does not present the end-of-life product, see spec.: http://www.digitalnow.com.au/product_pages/airstar2.html
[20] http://neuron2.net/dgmpgdec/dgmpgdec.html - developer site
[21] http://trans.sourceforge.net Transcriber 1.5.1 developer site
[22] http://www.foobar2000.org/ - developer site
[23] PLEVA, M., JUHAR, J., CIZMAR, A.: Multimedia Database Management for Annotators of the Metadata Content, *Proc. of RTT 2009*, Praha : CVUT, p. 3, 2009.
[24] http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/index.html – LDC recommendation
[25] NIST SCLITE scoring toolkit: http://www.itl.nist.gov/iad/mig/tools/
[26] POLLAK, P. et al.: SpeechDat(E)-Eastern Speech Databases, *Proc. of LREC 2000, XLDB satellite workshop,* Athens, Greece, pp. 20–25, 2000.
[27] ZGANK, A. et al.: The COST 278 Initiative – Crosslingual Speech Recognition with Large Telephone Database, *Proc. of LREC 2004*, Lisbon, May 2004, pp. 2107–2110.
[28] PLEVA, M.: Automatic Processing of Speech Data in Multimedia Databases (Automaticke Spracovanie Recovych Dat v Multimedialnych Databazach), *PhD Thesis (in Slovak)*, FEI TU of Kosice, p. 93, 2009.
[29] LINDBERG, B. et al.: A Noise Robust Multilingual Reference Recogniser Based on Speechdat (II), *Proc. of INTERSPEECH 2000*, Beijing, China, October 16–20, 2000, pp. 370–373, 2000.
[30] IVANECKY, J., NABELKOVA, M.: Phonetic transcription SAMPA and Slovak language (in Slovak), *Jazykovedny casopis,* vol. 53, pp. 81–95, 2002.