An Extension of the Slovak Broadcast News Corpus based on Semi-Automatic Annotation

Peter Viszlay, Ján Staš, Tomáš Koctúr, Martin Lojka, Jozef Juhár
Department of Electronics and Multimedia Communications
Faculty of Electrical Engineering and Informatics, Technical University of Košice
Park Komenského 13, 042 00 Košice, Slovak Republic
{peter.viszlay, jan.stas, tomas.koctur, martin.lojka, jozef.juhar}@tuke.sk

Abstract

In this paper, we introduce an extension of our previously released *TUKE-BNews-SK* corpus based on a semi-automatic annotation scheme. It firstly relies on the automatic transcription of the BN data performed by our Slovak large vocabulary continuous speech recognition system. The generated hypotheses are then manually corrected and completed by trained human annotators. The corpus is composed of 25 hours of fully-annotated spontaneous and prepared speech. In addition, we have acquired 900 hours of another BN data, part of which we plan to annotate semi-automatically. We present a preliminary corpus evaluation that gives very promising results.

Keywords: automatic subtitling, broadcast news, semi-automatic annotation, speech-to-text transcription

1. Motivation and Previous Work

The initial research related to Broadcast News (BN) transcription for the Slovak language started in recent years, when the BN data collection was performed. This activity resulted in the first Slovak *TUKE-BNews-SK* corpus, introduced in (Pleva and Juhár, 2014). It consists of 265 hours of spontaneous speech, dialogues and live coverage with different background conditions. It was manually annotated, because there were no other available resources supporting the annotation process at that time. It is still the only Slovak annotated corpus for building knowledge sources for the BN transcription task.

In this paper, we introduce a new extension of the *TUKE-BNews-SK* corpus, which stems from the current needs of our laboratory, and whose creation was partially initiated by demand in Slovak research. The reason for building the corpus lies in the fact that the modern and leading trends in building resources for large vocabulary continuous speech recognition (LVCSR) applications focus on fully-automatic annotation of speech (Wessel and Ney, 2005; Novotney et al., 2009), without any additional human annotation effort. This fact motivated us to follow this trend.

Unfortunately, we still do not have sufficient acoustic resources to perform automatic annotation, because the previous corpus did not provide enough data to build accurate acoustic models (AMs) for that purpose. With our increasing effort to build more precise and robust speech transcription system for error free speech transcription, the need for more data of high-quality become stronger and resulted recently in a cooperation with one of the Slovak commercial TV broadcasters. Moreover, with increasing demand driven by the growing population of deaf and hearing impaired people (approx. $5-7\%$ of Slovak population), we started recently with this broadcaster to support more TV broadcasting with Slovak closed captions generated automatically through modern speech recognition technologies.

The above mentioned facts really motivated us to create a new corpus, which would meet the conditions and requirements for building modern LVCSR systems suitable for automatic annotation purposes. In the first phase, the presented annotated corpus was considered as initial data intended to bootstrap our speech transcription system and to find its optimal configuration. The initial data were employed also for gender detection, speaker diarization and identification, for preparing the adaptation data for acoustic modeling and finally, for the preliminary evaluation in terms of word error rate (WER) for different BN categories. The corpus was considered to be sufficient after we had collected a reasonable amount of acoustic data with respect to the adequate speaker coverage. It finally amounted to 25 hours of fully-transcribed high quality multi-track acoustic data originally provided to us by the broadcaster. So far, we have collected 900 hours of raw, non-annotated acoustic data to be annotated, including primarily prime time broadcast and sports news, social programs and TV shows.

2. Corpus Design

There are several significant differences between the previously released corpus and its extension. The data we have collected are composed of multi-channel audio recordings (see Sec. 2.1.), whereas the *TUKE-BNews-SK* corpus gathers single-channel recordings. Each channel has a separated content (in-studio anchors, external reporters and interviewees). The audio data from the in-studio anchors were captured directly in the TV studio by lavalier microphones. The voices of the external reporters were captured by handheld microphones mixed with the signal from noise-cancelling camera microphone. For annotation purposes the source audio channels were simply joined together. On the other hand, the multi-channel mode permits to employ advanced techniques for precise speech processing, noise suppression and acoustic modeling.

Furthermore, the new speech database covers current events and hot topics in Slovakia, which is a suitable ground for domain-based modeling and topic extraction tasks. The current speaker inventory is expanded because of the presence of a number of new speakers. Some specific speakers remained the same and hence their database was successfully extended. Last, but not least, we cannot omit the fact

BN category	number of speakers		
	male	female	all
in-studio anchors	3	3	6
external reporters	33	19	52
known interviewees	157	59	216
unknown interviewees	1,350		1,350
TOTAL			1,624

Table 1: Speaker coverage in the new corpus

focus conditions	**duration** [hh:mm:ss]
F0 – prepared speech in studio	17:39:06
F1 – spontaneous speech in studio	03:37:47
F2 – prepared telephone speech (reduced-bandwidth)	00:05:24
F3 – speech with music in background (SNR$< 10dB$)	00:23:08
F4 – speech under degraded ac. conditions	02:31:52
F5 – speech performed by non-native speakers	00:01:39
Fx – combination of the focus conditions listed above (F1-F5)	00:23:39
TOTAL	24:42:35

Table 2: Data distribution across focus conditions

that speaking style has changed slightly during the last few years, which should be taken into account in acoustic and language modeling.

Apart from the annotated corpus, we have collected other related acoustic data to be annotated, which comprise the following three blocks:

1. **prime time evening TV News with Sports News** (330 hours) – covered by three in-studio anchors and many reporters mainly outside the studio;
2. **one-hour block of afternoon program** (191 hours) – it consists of Telephone Lottery contest, TV News with similar, but reduced content of prime time TV News, and social program, called "Reflex";
3. **the "Morning" TV shows** (380 hours) – live broadcast often from outside, where two anchors lead semi-prepared discussions with more than 10 guests about cooking, health, lifestyle, fashion, etc. There is a lot of background noise or music in the background.

2.1. Data Acquisition and Archiving

The source data are acquired as video tracks in MP4 container format and also as multi-channel audio tracks of sampling frequency $f_s = 48kHz$ in $24bit$ linear PCM (L24) format. Since the AMs were originally trained on speech data with $f_s = 16kHz$, the audio data have to be downsampled to equal frequency. We utilized a multi-channel speech recognition operated through a web user interface (Koctúr et al., 2015). For that purpose, we store the data in multimedia container Matroska[1] in separate audio streams to one file without audio/video compression (Staš et al., 2015).

2.2. Basic Corpus Statistics

The presented corpus consists overall of 60 hours of annotated content, including the silence and other malformed audio content. The useful part covers a total length of about 25 hours of fully-annotated speech, encompassing approximately $215k$ words. The corpus covers overall $1,624$ various speakers of different ages divided into 4 categories: in-studio anchors, external reporters, known annotated and unknown interviewees, according to the Table 2..

Intuitively, there are different speech styles, exactly covered by 76.70% of prepared and semi-prepared speech (anchors, reporters) and by 23.30% of spontaneous speech (interviewees). The speaking rate varies from 141.31 up to 185.59 words per minute (wpm). The average rate of the out of vocabulary (OOV) words is 2.58% and the average perplexity (PPL) is 436.28. The gender distribution across the corpus results in 60% (approx. 15.25 hours) for male and 40% (approx. 9.50 hours) for female speakers.

2.3. Semi-Automatic Corpus Annotation

The absence of acoustic resources needed for the automatic annotation led us to use a middle ground solution, which lies in semi-automatic annotation. This concept is closely related to lightly supervised acoustic model training, where a well-trained LVCSR system is firstly used for initial transcription (Lamel et al., 2002; Li et al., 2013). It is a well-established way to tackle the lack-of-data problem, especially for languages that do not belong to the group of major EU languages. Thus, we firstly used our server-based LVCSR system (Staš et al., 2015) to transcribe the speech as well as possible. Secondly, we employed a number of trained human annotators for the subsequent manual post-annotation at word level.

The speech recognition server is based on large vocabulary recognition engine Julius (Lee et al., 2001) that was modified to support multi-threaded parallel speech recognition and sharing acoustic and language models (LMs) among all instances for memory space saving purposes (Lojka et al., 2014). The acoustic model has been trained on gender-balanced acoustic databases that included:

- 250 hours of annotated speech recordings from *TUKE-BNews-SK* corpus (Pleva and Juhár, 2014);
- 250 hours of annotated speech recordings of judicial readings, read phonetically rich sentences, newspaper articles and spelled items, recorded in conference rooms (Rusko et al., 2014);
- 90 hours of annotated speech recordings realized in the main hall of the Slovak Parliament (Darjaa et al., 2011; Rusko et al., 2014);
- 80 hours of annotated speech recordings from TV shows named "Court Room" (Rusko et al., 2014).

The trigram Slovak LM has been created using the SRILM toolkit (Stolcke, 2002), restricted to the vocabulary size of $416k$ unique words and smoothed by the Witten-Bell algorithm. The LM adapted to the domain of broadcast news, has been trained on preprocessed and classified text corpora of more than $2,150M$ tokens contained in $120M$ Slovak sentences (Hládek et al., 2014; Staš and Juhár, 2015). The speech transcription system further employed

[1] https://www.matroska.org/

BN category	gender	number of utterances	number of words	speaking rate [wpm]	OOV rate [%]	PPL	WER [%]	
							AM	GD AM
in-studio	male	1,958	25,165	157.92	2.56	435.85	18.51	17.49
anchors	female	2,244	25,955	143.81	2.49	398.44	15.61	14.74
external	male	4,230	49,268	159.98	3.08	404.40	17.38	17.56
reporters	female	2,384	26,909	157.10	1.65	500.87	13.80	13.80
known	male	892	8,906	165.84	2.08	419.72	35.03	34.53
interviewees	female	245	2,514	155.66	2.07	315.33	31.70	30.51
unknown	male	2,613	25,958	161.46	2.91	494.33	39.19	38.60
interviewees	female	1,214	11,705	157.77	2.63	446.95	36.33	36.34
TOTAL		15,780	176,380	156.69	2.58	436.28	22.30	21.94

Table 3: Experimental results of the extended *TUKE-BNews-SK* corpus evaluation

gender detection, speaker diarization and identification, multi-channel speech segmentation, and multi-pass sequential speech recognition with hypothesis combination (Lojka and Juhár, 2014; Kiktová and Juhár, 2015; Staš et al., 2015). Since the transcribed acoustic data were not accurate (approx. 21.40% WER), the generated annotations were completed by annotators using the Transcriber tool (Barras et al., 2001) with respect to the corrections and by adding gender and speaker labels, acoustic condition labels and focus conditions. The background sounds, such as music, background speech, different noises, telephone speech, etc., were also manually annotated (see Tab. 2.), because we are planning to build acoustic models that are robust towards these noises through multi-condition training. The annotations permit full speaker-adaptive training.

In the second phase, we are planning to append the fully-annotated data from that corpus to the current training data in order to retrain the present acoustic and language models. The retrained sources will be used to perform further automated transcription and manual correction of a different set of acquired BN data. Consequently, the new acoustic data will be appended again to the already expanded training data and the process of LVCSR system retraining and semi-automatic annotation will iteratively proceed until a sufficient amount is reached (usually hundreds of hours). Naturally, the amount of training data will grow rapidly, and the transcription accuracy at each iteration should consistently increase. From this point, we hope that we will be able to move to the automatic annotation.

3. Experimental Results

We report here the preliminary results (see Table 2.2.) obtained by recognizing the annotated BN data employing our available Slovak LVCSR system and evaluating them using the corresponding reference annotations according to four gender-specific BN categories (in-studio anchors, external reporters, known and unknown interviewees). The results are divided into six categories: number of utterances, number of words, speaking rate, OOV rate, perplexity (PPL) and finally, WER values, computed in the standard way. The BN categories were evaluated separately by general AM and by two gender-dependent models (GD AM) that were trained using only the gender-specific training data.

The results, listed in the Table 2.2., are very promising at the first glance. The WER values vary from 13.80% to 38.60% according to the BN category. As was expected, the lowest error rates are achieved for in-studio anchors and for the external reporters whose speech is well recognizable. This is partly due to the fact that the speech is prepared or semi-prepared and the sound quality is usually good. It is worth noting that the error rates correspond to the OOV rate. It is obvious that the lowest WER (13.80%) is achieved at the lowest OOV rate (1.65). A growing trend of OOV–WER rates is observable for the other BN categories. This is caused by higher number of unknown and new words in the utterances that absent in the dictionary.

The WER values of the last two categories are considerably higher (31.70%–39.19%). This is caused by the fact that non-native speakers frequently occur in the utterances or the speech is mostly spontaneous and moreover, the acoustic conditions are often degraded by background or traffic noise or by other sounds (music, shouting, sirens, wind, etc.). Note that these acoustic conditions have not been included in the acoustic modeling yet.

Regarding the gender-dependent evaluation, we can state that it can bring a further reduction of WER (except three cases) in the range from -0.50% to -1.19% absolutely. We have carried out also 58 speaker-dependent evaluations for in-studio anchors and external reporters (see Table 2.) that are not listed in the Table 2.2.. More precisely, the speaker-level WER values vary roughly from 6.15% to approximately 46.32% depending on the type of speaker, its gender, speaking and OOV rates, acoustic conditions, etc.

An other view to the corpus evaluation provides the Table 3. that summarizes the results according to the focus conditions (see Table 2.). The table demonstrates that the lowest WER is obtained for the category of prepared speech in studio (F0) that has the most acoustic representations (approx. 66%). Acceptable results are also achieved for the spontaneous speech in studio (F1). The error rates of higher-order focus conditions (F2–FX) are markedly higher, because the AMs did not exactly match the acoustic conditions or the type of speech present in the recognized segments.

Based on the previous analysis, we can conclude that the presented preliminary results are promising, but of course, there is still room for improvements, regarding various issues, such as spontaneous speech recognition, speaker-adaptive training, adaptive acoustic model retraining, multicondition training, dynamic vocabulary adaptation, etc.

focus conditions	**number of utterances**	**number of words**	**PPL**	**WER [%]**
F0	10,404	123,477	436.76	15.85
F1	2,680	26,217	461.51	31.86
F2	46	636	333.05	52.67
F3	272	2,680	553.66	37.01
F4	2,053	19,882	438.82	46.58
F5	30	242	416.71	73.97
FX	295	3,246	487.29	41.25
TOTAL	15,780	176,380	436.28	21.94

Table 4: Experimental results for different focus conditions

4. Conclusions and Future Intentions

In this paper, we introduced a new extension of our previously released *TUKE-BNews-SK* corpus. We were motivated by the modern trends in corpora design and we employed a semi-automatic annotation procedure to generate error-free transcriptions. Our current intention is focused on the semi-automatic acoustic data annotation, so far. We expect and hope that we will be able in the near future to move to fully-automatic annotation of any amount and kind of new data without the need for human annotation effort.

Our other serious interest is focused on the DNN-based LVCSR for Slovak using Kaldi (Povey et al., 2011). Therefore, we would like to replace the current speech recognition engine Julius by the WFST-based Kaldi engine. We are also interested in real-time BN subtitling and focus on improving the subtitling performance to generate closed captions for deaf and hearing impaired people.

5. Acknowledgements

The research presented in this paper was partially supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project VEGA 1/0075/15 (50%) and the Research and Development Operational Programme funded by the ERDF under the project implementation University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS code: 26220220182 (50%).

6. Bibliographical 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*, 33(1–2):5–22.

Darjaa, S., Cerňak, M., Trnka, M., Rusko, M., and Sabo, R. (2011). Effective triphone mapping for acoustic modeling in speech recognition. In *Proc. of INTERSPEECH 2011*, pages 1717–1720, Florence, Italy.

Hládek, D., Ondáš, S., and Staš, J. (2014). Online natural language processing of the Slovak language. In *Proc. of the 5th IEEE Conference on Cognitive Infocommunications, CogInfoCom 2014*, pages 315–316, Vietri sul Mare, Italy.

Kiktová, E. and Juhár, J. (2015). Comparison of diarization tools for building speaker database. *Advances in Electrical and Electronic Engineering*, 13(4):314–319.

Koctúr, T., Pleva, M., and Juhár, J. (2015). Interface for smart audiovisual data archive. In *Proc. of 25th International Conference RADIOELEKTRONIKA 2015*, pages 292–294, Pardubice, Czech Republic.

Lamel, L., Gauvain, J.-L., and Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16(1):115–129.

Lee, A., Kawahara, T., and Shikano, K. (2001). Julius - An open source real-time large vocabulary recognition engine. In *Proc. of EUROSPEECH 2001*, pages 1691–1694, Aalborg, Denmark.

Li, X., Pang, Z., and Wu, X. (2013). Lightly supervised acoustic model training for Mandarin continuous speech recognition. In *Intelligent Science and Intelligent Data Engineering*, LNCS 7751, pages 727–734. Springer Berlin, Heidelberg.

Lojka, M. and Juhár, J. (2014). Hypothesis combination for Slovak dictation speech recognition. In *Proc. of the 56th International Symposium ELMAR 2014*, pages 1–4, Zadar, Croatia.

Lojka, M., Ondáš, S., Pleva, M., and Juhár, J. (2014). Multi-thread parallel speech recognition for mobile applications. *Journal of Electrical and Electronics Engineering*, 7(1):81–86.

Novotney, S., Schwartz, R., and Ma, J. (2009). Unsupervised acoustic and language model training with small amounts of labelled data. In *Proc. of ICASSP 2009*, pages 4297–4300, Taipei, Taiwan.

Pleva, M. and Juhár, J. (2014). TUKE-BNews-SK: Slovak broadcast news corpus construction and evaluation. In *Proc. of LREC 2014*, pages 1709–1713, Reykjavik, Iceland.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proc. of ASRU 2011*, pages 1–4, Big Island, Hawaii, USA.

Rusko, M., Juhár, J., Trnka, M., Staš, J., Darjaa, S., Hládek, D., Sabo, R., Pleva, M., Ritomský, M., and Lojka, M. (2014). Slovak automatic dictation system for judicial domain. In *Human Language Technology Challenges for Computer Science and Linguistics*, LNAI 8387, pages 16–27. Springer, Switzerland.

Staš, J. and Juhár, J. (2015). Modeling of Slovak language for Broadcast News transcription. *Journal of Electrical and Electronics Engineering*, 8(2):43–46.

Staš, J., Viszlay, P., Lojka, M., Koctúr, T., Hládek, D., Kiktová, E., Pleva, M., and Juhár, J. (2015). Automatic subtitling system for transcription, archiving and indexing of Slovak audiovisual recordings. In *Proc. of the 7th Language & Technology Conference, LTC 2015*, pages 186–191, Poznań, Poland.

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. In *Proc. of ICSLP 2002*, pages 901–904, Denver, Colorado, USA.

Wessel, F. and Ney, H. (2005). Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13:23–31.