TEDXSK AND JUMPSK: A NEW SLOVAK SPEECH RECOGNITION DEDICATED CORPUS

JÁN STAŠ – DANIEL HLÁDEK – PETER VISZLAY – TOMÁŠ KOCTÚR
Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovakia

STAŠ, Ján – HLÁDEK, Daniel – VISZLAY, Peter – KOCTÚR, Tomáš: TEDxSK and JumpSK: A New Slovak Speech Recognition Dedicated Corpus. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 346 – 354.

Abstract: This paper describes a new Slovak speech recognition dedicated corpus built from TEDx talks and Jump Slovakia lectures. The proposed speech database consists of 220 talks and lectures in total duration of about 58 hours. Annotated speech database was generated automatically in an unsupervised manner by using acoustic speech segmentation based on principal component analysis and automatic speech transcription using two complementary speech recognition systems. The evaluation data consisting of 50 manually annotated talks and lectures in total duration of about 12 hours, has been created for evaluation of the quality of Slovak speech recognition. By unsupervised automatic annotation of TEDx talks and Jump Slovakia lectures we have obtained 21.26% of new speech segments with approximately 9.44% word error rate, suitable for retraining or adaptation of acoustic models trained beforehand.

Keywords: automatic annotation, speech recognition, speech corpus

1 INTRODUCTION

The development of more advanced and more precise large vocabulary continuous speech recognition (LVCSR) system requires huge amount of data for estimation of statistical parameters of acoustic and language models to cover the most possible real situations that usually occur in spontaneous speech. The complexity of speech recognition is mainly influenced by the speaker characteristics and speaking style. Robust acoustic models require phonetically rich and gender-balanced speech corpora that contain from hundreds to thousands of hours of annotated speech recordings.

Manual speech transcription and annotation of such amount of data requires much time and effort, as well as considerable amount of funds. Conventional transcription and annotation methodology requires training of the professional annotators on transcription guidelines, which lasts from a few hours to several weeks. Typical manual transcription speeds of spontaneous or conversational speech lasts around 7 to 12 times real-time, due to its complexity. The transcription and annotation of non-native speech is an even more difficult, slow and laborious process [1].

If even a few hours of manually annotated speech utterances is available, then it is possible to develop, using the latest approaches, principles and methods, a comprehensive speech transcription system for automatic annotation and creation of a new speech database that can be used for re-estimating selected parameters of an existing acoustic models or their adaptation to the characteristics of the given speaker.

DOI 10.1515/jazcas-2017-0044

Publicly available online spoken language resources are preferable because there are problems with obtaining licence agreement from source data providers. One such resource is the database of TED talks (Technology, Entertainment, Design) that promote "*ideas worth sharing*". They become a good online available resource for creation of speech databases for the number of languages, which are under-resourced because of their thematic, stylistic and natural richness.

One of the best known and widely used automatic speech recognition dedicated corpus is TED-LIUM [2] that consists of 1495 automatically annotated English talks. The initial speech recognition was performed using five-pass ASR system based on the open-source CMU Sphinx framework [3] with acoustic and language model adaptation, speaker adaptive training, re-computing the linguistic scores from updated word-graphs with 4-gram language model and algorithm for hypothesis re-scoring at different stages. The official speech recognition results discussed at the IWSLT 2011 evaluation campaign reached 17.40% word error rate (WER) in average.

Otherwise, one of recently proposed and automatically annotated spoken language resource built from TED talks is the SI TEDx-UM speech database [4]. It contains 242 talks in the Slovenian language in total duration of about 54 hours. The efficiency of unsupervised transcription was evaluated using the UMB Broadcast News speech recognition system and reached 50.70% WER in average.

Several previous works and research activities have been reported on enhancing the efficiency of automatic speech transcription and unsupervised annotation of speech corpora based on TED talks by using robust acoustic and language modeling [5-8].

A number of algorithms have been proposed for acoustic model adaptation in automatic transcription of TED talks based on discriminative training criteria, maximum a posteriori (MAP) estimation, maximum likelihood linear regression (MLLR), feature space adaptation (FSA), vocal tract length normalization (VTLN) [9], speaker adaptive training (SAT) [10], or statistical modeling with deep neural networks (DNN) [11].

Moreover, the problem of frequently appearing errors in automatic transcription of lecture speech was eliminated in [12] by correction of colloquial expressions, deletion of fillers and insertion of periods using statistical post-processing techniques. Authors in [13] and [14] explore output recognition hypotheses and effectiveness of supervised and unsupervised adaptation with varying amounts of user-provided transcripts to tune the language model parameters on a lecture transcription task in English.

This paper presents a new spoken language resource built from Slovak TEDx talks and Jump Slovakia lectures annotated automatically in an unsupervised manner using two complementary LVCSR systems with using filtration of output hypotheses with minimal amount of errors. The reason for building the corpus lies in the fact that modern and leading trends in building resources for LVCSR applications focus on fully-automatic annotation of speech, without any additional human effort [2][4] [10][15]. The database of speech recordings and their transcripts will be publicly available from the end of june 2017 at the web page of our laboratory[1].

[1] `http://nlp.web.tuke.sk/pages/tedx`

2 CONTENT AND STRUCTURE OF THE SPEECH CORPUS

The goal of this research is to build a new automatically annotated speech database with the best possible quality and one or at least two speakers per talk. Source data consisting of Slovak TEDx talks[2] and Jump Slovakia[3] lectures were gathered from official YouTube channels. Foreign-language lectures and low-quality speech recordings were removed from the list of about 300 Slovak audiovisual recordings obtained from ten different events, publicly available between years 2010 and 2016.

event	number of lectures	number of speakers	males	females	duration	for male gender	for female gender
TEDx Bratislava	57	61	42	19	13:03:55	09:02:35	04:01:20
TEDx Kežmarok	9	10	6	4	02:48:06	01:59:18	00:48:48
TEDx Košice	30	30	24	6	08:50:03	07:24:35	01:25:28
TEDx Nitra	14	14	12	2	04:13:37	03:33:07	00:40:30
TEDx Prešov	17	17	11	6	05:57:31	04:07:32	01:49:59
TEDx Trenčín	24	25	14	11	05:50:43	03:36:40	02:14:03
TEDx Trnava	9	9	6	3	02:21:53	01:42:20	00:39:33
TEDxYouth Bratislava	20	20	15	5	05:36:39	04:06:05	01:30:24
TEDxYouth Žilina	6	6	4	2	01:41:34	01:06:59	00:34:35
Jump Slovensko	34	35	20	15	07:25:35	04:12:44	03:14:51
together	**220**	**227**	**154**	**73**	**57:51:36**	**40:51:55**	**16:59:41**

Tab. 1. Structure of the speech corpus of TEDx talks and Jump Slovakia lectures

2.1 Corpus Design

A total of 220 talks and lectures in the Slovak language were selected for automatic segmentation and unsupervised transcription using our proposed system architecture based on two complementary Slovak LVCSR systems. Audiovisual recordings downloaded from the official YouTube channels were encoded in H.264 video format. The captured audio stream was encoded in MPEG AAC format. Each recording has been converted into WAV audio and down-sampled to 16kHz/16bit PCM mono audio using SoX tool[4] to be compatible with acoustic models used in our LVCSR system.

2.2 Corpus Statistics

The presented speech corpus consists of about 58 hours, including silence and other malformed audio content. The useful part covers a total duration of about 55 hours. The speech corpus contains 227 unique speakers of both genders with 154 males and 73 females. Approximately 30% of the database is build of samples of female voices.

Manually annotated part of the speech corpus covers 50 randomly selected Slovak TEDx talks in total duration of about 12 hours. The speaking rate in this part

[2] https://www.youtube.com/user/TEDxTalks
[3] https://www.youtube.com/user/jumpslovensko
[4] http://sox.sourceforge.net/

varies from 115.53 up to 256.32 words per minute (*wpm*). The average rate of out-of-vocabulary (OOV) words is 3.23% and the average language model perplexity is 508.40. The detailed description about the number of lectures, number of speakers and total duration of speech in the presented corpus of 220 Slovak TEDx talks and Jump Slovakia lectures is summarized in the Table 1.

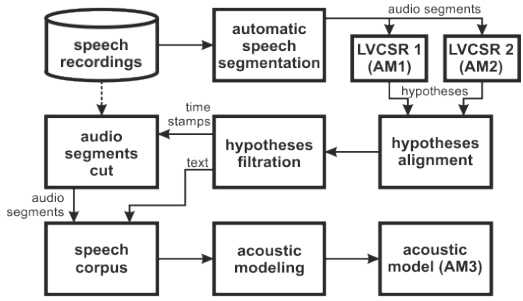

Fig. 1. Automatic segmentation and transcription of speech recordings using two complementary large vocabulary continuous speech recognition systems in the Slovak language

2.3 Characteristics of the Speech in the Corpus

During about 15-minute talk (lecture), speakers are often non-native, have a strong accent, and sometimes, are not fluent. Despite the fact that speaking style of a speaker being in general planned, spontaneous speech occurs more frequently. There are some differences among speakers in their grammar, articulation or speaking style with frequent errors (e.g. filled pauses, sentence restarts, phrase modifiers, repetitions, or mistakes). Although speech recordings are usually realized with close-talk, lapel or goose-neck microphones, the signal often contains some noise from the auditorium and from the speaker itself. Therefore, lecture speech transcription is a difficult task, both from the acoustic and linguistic point of view, due to the many hesitation fillers that occur in spontaneous speech, different and varying speaking rates, mixed topics and speaking style, or combining colloquial expressions with formal jargon [7]. Furthermore, a new speech corpus covers current events and hot topics in Slovakia, which is suitable ground for domain-based modeling and text summarization tasks.

3 AUTOMATIC SEGMENTATION AND ANNOTATION OF THE SPEECH CORPUS

The proposed approach for automatic segmentation and unsupervised speech transcription and annotation of the presented corpus of TEDx talks and Jump Slovakia lectures using two complementary LVCSR systems is depicted on Fig. 1. A brief review of the main building blocks of this system architecture will be described in the following sections.

3.1 Automatic Speech Segmentation

In general, it is possible to transcribe continuous audio stream without any segmentation, but the computation cost of the decoding may take a very long time.

Therefore, automatic speech segmentation is usually applied to speed up the speech recognition process and to improve the overall performance by identifying and handling the specific parts in the recognized speech (gender- and speaker-specific segments, speaker-change boundaries, different acoustic conditions, non-speech events, etc.).

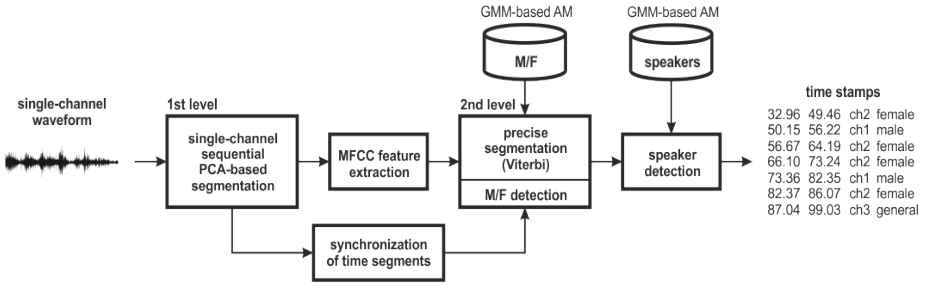

Fig. 2. Automatic speech segmentation

The proposed speech segmentation is able to process any kind of single-channel audio recordings (e.g. talks, lectures, discussions, broadcast news, etc.). The gender detection can be performed using the default gender-dependent acoustic models. The detection rate will be satisfactory, because the acoustic models were trained on a sufficient amount of acoustic representations for each gender.

The speaker-dependent segmentation is not supported implicitly, if the single-channel waveform contains voices of unknown speakers that were not included in the training data. On the other hand, there is a possibility to train a new speaker-dependent AM, if the recognized audio provides a sufficient amount of speaker examples.

The proposed system architecture employs two-level fully-automatic speech segmentation, depicted on Fig. 2.

At first level, the silence discrimination is performed by our proposed voice activity detection (VAD) algorithm [16]. In order to determine VAD labels, the waveform is processed in the time domain by overlapping blocks extracted by rectangular window with length of 25ms and 10ms frame step. After re-arranging samples into sample matrix, the time domain principal component analysis (PCA) is applied to each block. After that, N eigenvalues are computed for each block, where N is the dimension of the PCA space. The eigenvalues are used to determine the nature of the i-th segment (voice or silence). Finally, the VAD coefficients are smoothed by applying a sliding average window.

The second level uses only the speech active segments and it employs the Viterbi algorithm for precise gender- or speaker-dependent segmentation. In other words, gender- and speaker-dependent recognizers are run to detect and locate gender- and speaker-change points to enable these regions to be split into shorter segments. At this stage, time stamps are generated with gender labels and if needed, they can be extended with speaker labels. The overall segmentation requires a time synchronization between the first and second level due to eliminated silent parts at the first level.

3.2 Automatic Speech Transcription

For initial experiments with unsupervised transcription, annotation and acquisition of large speech databases we have created a new speech recognition system architecture based on the complementarity of two Slovak LVCSR systems (see Fig. 1, LVCSR 1 and LVCSR 2) [1], [17].

The Slovak LVCSR system uses an open-source recognition engine Julius [18] that was modified to support multi-threaded parallel speech recognition and sharing acoustic and language models among all instances for memory space saving purposes. For supporting the actual speech recognition with different configurations, the speech recognition server was created and is capable of parallel speech recognition supporting different configurations at the same time [19].

Complementarity of the LVCSR systems was achieved by using two acoustic models trained on the different training sets. The first acoustic model (AM 1) was trained on 320 hours of manually annotated speech recordings of judicial readings and parliament proceedings [20]. The second model (AM 2) was trained on a database of 330 hours of manually annotated speech recordings acquired from the main broadcast news [21] and Court TV shows with a high degree of spontaneity [1], [21]. Both acoustic models (AM 1 and AM 2) were generated from feature vectors with standard dimension of 39 mel-frequency cepstral coefficients, along with delta and acceleration coefficients and cepstral mean normalization enabled. The triphone context-dependent acoustic models are based on hidden Markov models (HMMs) with 32 Gaussian mixtures. The training sets also involve models of silence, short pause and additional noise events for filled pauses and prolongations. A typical tree-based state tying for HMMs has been replaced by the effective triphone mapping algorithm [22].

The proposed LVCSR system uses a trigram model of the Slovak language created by the SRILM Toolkit [23], restricted to the vocabulary size of about 500*k* unique words and smoothed with the Witten-Bell algorithm [24]. Language model has been trained on preprocessed web-based corpora of Slovak written texts of more than 2,150M tokens contained in 120M of sentences [25].

The process of speech recognition was enhanced using *N*-best output hypotheses rescoring with the ROVER algorithm [26], slightly modified to our needs to include confidence measure score context between words into consideration [27].

3.3 Filtration of Output Hypotheses

After transcription of segmented speech recordings, the output hypotheses from both complementary recognition systems (LVCSR 1 and LVCSR 2) are time aligned and compared. In the next step, overlapping transcribed speech segments obtained from aligned output hypotheses are filtered out. Proposed approach for filtration output hypotheses takes maximum time delay from the start and end of the speech segment, minimum number of equal words in aligned hypotheses and confidence measure score (CMS) value into account. Output of the process of filtration are short automatically annotated speech segments [19].

In this research, the parameters for filtration of output hypotheses were empirically set to 20ms maximum time delay from the start and end of each speech

segment, minimum number of equal words in aligned hypotheses was set to 3 words, and the threshold value for confidence measure score varies from 0 to 0.75.

data set	actual duration	duration after segmentation	setting 1 ~ 13.57% WER	setting 2 ~ 9.44% WER	setting 3 ~ 4.94% WER
	amount of gathered data [hh:mm:ss]				
eval	12:26:07	11:50:37	05:39:30	02:47:35	00:39:43
dev	45:25:29	43:13:12	19:37:41	08:54:47	02:01:04
eval+dev	57:51:36	55:03:49	25:17:11	11:42:22	02:40:47
	amount of gathered data [%]				
eval		95.24	47.78	23.58	5.59
dev		95.15	45.41	20.62	4.67
eval + dev		95.17	45.92	21.26	4.87

Tab. 2. Amount of gathered data

4 EVALUATION

In the first step of creation of a new spoken language resource of Slovak TEDx talks and Jump Slovakia lectures we divided speech corpus into two parts – evaluation and development data set. The evaluation data, in total duration of 12 hours, was annotated manually on the word level by professional annotators. This set was used for evaluation of the transcription accuracy in various settings (see Table 2, setting 1 to 3). These settings of the systems were selected for the best quality of the transcription (setting 1) and for the biggest amount of annotated data (setting 3). Setting 2 is a trade-off between quality and quantity of automatic speech transcription.

Values obtained from the settings (1-3) were used for automatic transcription of the remaining (development) part of the corpus. The total amount of gained data after automatic transcription is summarized in Table 2.

The Table 2 shows that it is possible to get 45.92% of a new fully-automatic annotated speech segments from the total length and amount of speech recordings with approximately 13.57% WER. These automatically annotated speech segments can be directly used for re-estimation of the parameters of existing acoustic models or for their adaptation. Similarly, 21.26% of new fully-automatic annotated speech segments can be gained with approximately 9.44% WER and about 4.94% WER can bring only 4.87% of annotations from the total amount of 58 hours.

5 CONCLUSION

In this paper we introduced a new speech recognition dedicated corpus built from Slovak TEDx talks and Jump Slovakia lectures. We were motivated by the modern trends in corpora design based on fully-automatic annotation procedure to generate error-free transcripts. We expect that we will be able in the immediate future to move fully-automatic annotation of any kind of new data without the need for human annotation effort.

In the further research, we want to focus on eliminating common recognition errors by introducing unsupervised language model adaptation to the current topic and specific speaker speaking style and statistical modeling of most frequent hesitation fillers in spontaneous speech for improving system performance and automatic transcription and annotation of large acoustic corpora of the spoken Slovak language [28]. Also, 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.

ACKNOWLEDGEMENTS

The research in this paper was supported by the Faculty of Electrical Engineering and Informatics at the Technical University of Košice under the project FEI-2015-30, by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project VEGA 1/0511/17, and by the Slovak Research and Development Agency under the applied research project APVV-15-0517.

References

[1] Koctúr, T., Juhár, J., Viszlay, P., Staš, J., and Lojka, M. (2016). Unsupervised speech transcription and alignment based on two complementary ASR systems. In *Proceedings of RADIOELEKTRONIKA 2016*, pages 358–362, Košice, Slovakia.

[2] Rosseau, A., Deléglise, P., and Estève, Y. (2012). TED-LIUM: An automatic speech recognition dedicated corpus. In *Proceedings of LREC 2012*, pages 125–129, Istanbul, Turkey.

[3] Deléglise, P., Estève, Y., Meignier, S., and Merlin, T. (2009). Improvements to the LIUM French ASR system based on CMU Sphinx: What helps to significantly reduce the word error rate? In *Proceedings of INTERSPEECH 2009*, pages 2123–2126, Brighton, UK.

[4] Žgank, A., Maučec, M. S., Verdonik, D. (2016). The SI TEDx-UM speech database: A new Slovenian spoken language resource. In *Proceedings of LREC 2016*, pages 4670–4673, Portorož, Slovenia.

[5] Rosseau, A., Deléglise, P., and Estève, Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of LREC 2014*, pages 3935–3939, Reykjavik, Iceland.

[6] Leeuwis, E., Federico, M., and Cettolo, M. (2003). Language modeling and transcription of the TED corpus lectures. In *Proceedings of ICASSP 2003*, pages 232–235, Hong Kong, China.

[7] Cettolo, M., Brugnara, F. and Federico, M. (2004). Advances in the automatic transcription of lectures. In *Proceedings of ICASSP 2004*, pages 769–772, Montreal, Canada.

[8] Niesler, T. and Willet, D. (2002). Unsupervised language model adaptation for lecture speech transcription. In *Proceedings of ICSLP 2002*, pages 1413–1416, Denver, Colorado, USA.

[9] Wölfel, M. and Berger, S. (2005). *The ISL baseline lecture transcription system for the TED corpus*. Tech. Rep., Karlsruhe University, Germany.

[10] Naptali, W. and Kawahara, T. (2012). Automatic transcription of TED talks. In *Proceedings of the 6th Spoken Document Processing Workshop, SDPWS 2012*, Toyohashi, Japan.

[11] Bell, P., Yamamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, Ch., and Renals, S. (2013). A lecture transcription system combining neural network acoustic and language models. In *Proceedings of INTERSPEECH 2013*, pages 3081–3091, Lyon, France.

[12] Nanjo, H., Shitaoka, K., and Kawahara, T. (2003). Automatic transformation of lecture transcription into document style using statistical framework. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, SSPR 2003*, Tokyo, Japan.

[13] Hsu, B.-J. and Glass, J. (2009). Language model parameter estimation using user transcriptions. In *Proceedings of ICASSP 2009*, pages 4805–4808, Taipei, Taiwan.

[14] Akita, Y., Watanabe, M., and Kawahara, T. (2012). Automatic transcription of lecture speech using language model based on speaking-style transformation of proceedings texts. In *Proceedings of INTERSPEECH 2012*, pages 2326–2329, Portland, Oregon, USA.

[15] Viszlay, P., Staš, J., Koctúr, T., Lojka, M., and Juhár, J. (2016). An extension of the Slovak broadcast news corpus based on semi-automatic annotation. In *Proceedings of LREC 2016*, pages 4684–4687, Portorož, Slovenia.

[16] Vavrek, J., Viszlay, P., Kiktová, E., Lojka, M., Juhár, J., and Čižmár, A. (2014). Query-by-example retrieval via fast sequential dynamic time warping algorithm. In *Proceedings of the 37th International Conference on Telecommunications and Signal Processing, TSP 2014*, pages 453–457, Berlin, Germany.

[17] 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 *Proceedings of the 7th Language & Technology Conference, LTC 2015*, pages 186–191, Poznań, Poland.

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

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

[20] Rusko, M., Juhár, J., Trnka, M., Staš, J., Darjaa, S., Hládek, D., Sabo, R., Pleva, M., Ritomský, M., and Ondáš, S. (2016). Advances in the Slovak judicial domain dictation system. In Vertulani, Z., Uszkoreit, H., and Kubis, M., editors, *Human Language Technology: Challenges for Computer Science and Linguistics*, LNAI 9561, pages 55–67, Springer International Publishing Switzerland.

[21] Koctúr, T., Staš, J., and Juhár, J. (2016). Unsupervised acoustic corpora building based on variable confidence measure thresholding. In *Proceedings of the 58th International Symposium ELMAR 2016*, pages 31–34, Zadar, Croatia.

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

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

[24] Staš, J. and Juhár, J. (2015). Modeling of the Slovak language for broadcast news transcription. *Journal of Electrical and Electronics Engineering*, 8(2):43–46.

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

[26] Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of ASRU 1997*, pages 347–352, Santa Barbara, CA, USA.

[27] Lojka, M. and Juhár, J. (2014). Hypothesis combination for Slovak dictation speech recognition. In *Proceedings of the 56th International Symposium ELMAR 2014*, pages 43–46, Zadar, Croatia.

[28] Staš, J., Hládek, D, and Juhár, J. (2016). Adding filled pauses and disfluent events into language models for speech recognition. In *Proceedings of the 7th IEEE International Conference on Cognitive InfoCommunications*, *CogInfoCom 2016*, Wroclaw, Poland.