The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 325-333

Building of children speech corpus for improving automatic subtitling services

Matus Pleva, Stanislav Ondas, Daniel Hládek, Jozef Juhar, Ján Staš
Department of Electronics and Multimedia Communications
Technical University of Kosice, Slovakia
matus.pleva@tuke.sk stanislav.ondas@tuke.sk daniel.hladek@tuke.sk
jozef.juhar@tuke.sk jan.stas@tuke.sk

Yuan-Fu Liao
Department of Electronic Engineering
National Taipei University of Technology
yfliao@mail.ntut.edu.tw

Abstract

This paper describes the development and first evaluation of the new Slovak children speech audio corpus for improving the automatic broadcast news subtitling engine developed on the Technical University of Kosice in cooperation with the Slovak Academy of Sciences. The current automatic speech recognition (ASR) systems are reliable for a clean, prepared speech of adults with not very long pause inside the sentences. For speech recognition of children's, it is still a challenge from different reasons. They use much slang, and diminutive words, undeveloped pronunciation, shorter vocal tract (different speech parameters), the sentence syntax is different. The paper presents the results of the children speech automatic recognition from the system built for broadcast news transcription.

Keywords: automatic speech recognition, children speech, audio corpus, annotation, database design.

1. Introduction

The speech technology has significant potential, currently it has growing interest among children and technically enthusiastic people [1]. The International Speech and Communication Association (ISCA) has a Special Interest Group (SIG) for Child Computer Interaction (CHILD) [2] and is organizing a special Workshop on Child Computer Interaction - WOCCI and last years also Language Teaching, Learning and Technology - LTLT. This year a special

session on the prestigious Interspeech conference will be held in September 2019 in Graz called Spoken Language Processing for Children's Speech [3].

The development of children's speech corpora for different languages is in progress [4] (British English, German and Swedish), [5] (non-native English), [6] Chinese, [7] Cantonese, [8] Jamaican English, [9] interactive emotional children speech and many others. For European union also small European languages are essential for electronic communication, so we decided to start the building of Slovak children speech corpus for improvement of the Slovak automatic speech recognition engines already built [10, 11]. Of course, the speech parameters different for children speech because of different vocal tract sizes [12], and they are many algorithms (Vocal-Tract Length Normalization - VTLN) how to handle it [13]. For children's speech, the formant frequencies are higher, the speech rate is slower or higher than in adult speech, and the language contains more home slang, garbled and imaginary words.

The Slovak language belongs to a group of Slavic languages, which are typical of inflection and free word order, which means it is morphologically rich and uses a very large vocabulary [10, 14]. These features make the Slovak automatic speech recognition task very complicated, and a large amount of data is required for automatic large vocabulary spontaneous speech recognition [14].

This article describes the first step, the collection of the first data, manual annotation, and testing of the current ASR system with children and adult speech recordings.

2. Building the database

For children's speech, there are very few freely available recordings on the Internet, especially in the form suitable for speech recognition system acoustic model training. We decided to use the TV series' recordings. There were several problems when using TV series recordings. The vast majority of segments are tinged with music, which would not matter if we were trying to build a model that would recognize where the sound begins and ends. However, when it comes to recognizing children's speech, it can cause distortions that will be undesirable for our purpose, and our results will be affected to some extent [15].

The main problem with a database suitable for acoustic models training is the resources needed for quality data annotation. This task is very time-consuming, and another reason is the lack of publicly available data, and therefore, our database is of a more modest size [15].

The database of children's recordings is made up of segments of children's speech from the television TV series of commercial Slovak broadcasters Markíza and JoJ. Specifically, they are the TV series Daddy (Oteckovia), broadcast on Markíza since early 2018, and Holidays (Prázdniny) from JoJ, the first part aired January 18, 2017.

The recordings were downloaded from premium archives of the TV broadcasters in Full HD. We cut the utterances with children speech out from the .MP4 recording and merged the parts without background music. The audio codec used in original file was AAC LC (Advanced Audio Coding - Low Complexity profile) with 48kHz 257.05 kbps Stereo settings.

Then the WAV file was exported in 48kHz Stereo PCM format and annotated with the Transcriber [16] application (Figure 1.). The collected database statistic is summarized in the Table 1.

Table 1. Database statistics

TV Series_Episode (Date)	Lenght [minutes]	Number of words
Oteckovia_E1(1.1.2018)	2:59	298
Oteckovia_E2(2.1.2018)	5:15	550
Oteckovia_E3(3.1.2018)	4:35	601
Oteckovia_E4(4.1.2018)	2:56	364
Oteckovia_E5(5.1.2018)	4:14	510
Oteckovia_E6(8.1.2018)	3:49	519
Oteckovia_E7(9.1.2018)	4:57	650
Prazdniny_E1(18.1.2017)	4:23	513
Prazdniny_E2(25.1.2017)	7:58	739
Total	**41:01**	**4744**

3. Transcription process

In our database, we have annotated the age, real names and surnames of publicly known children actors, so that we can see how the system performs with different ages of children (Figure 1.).

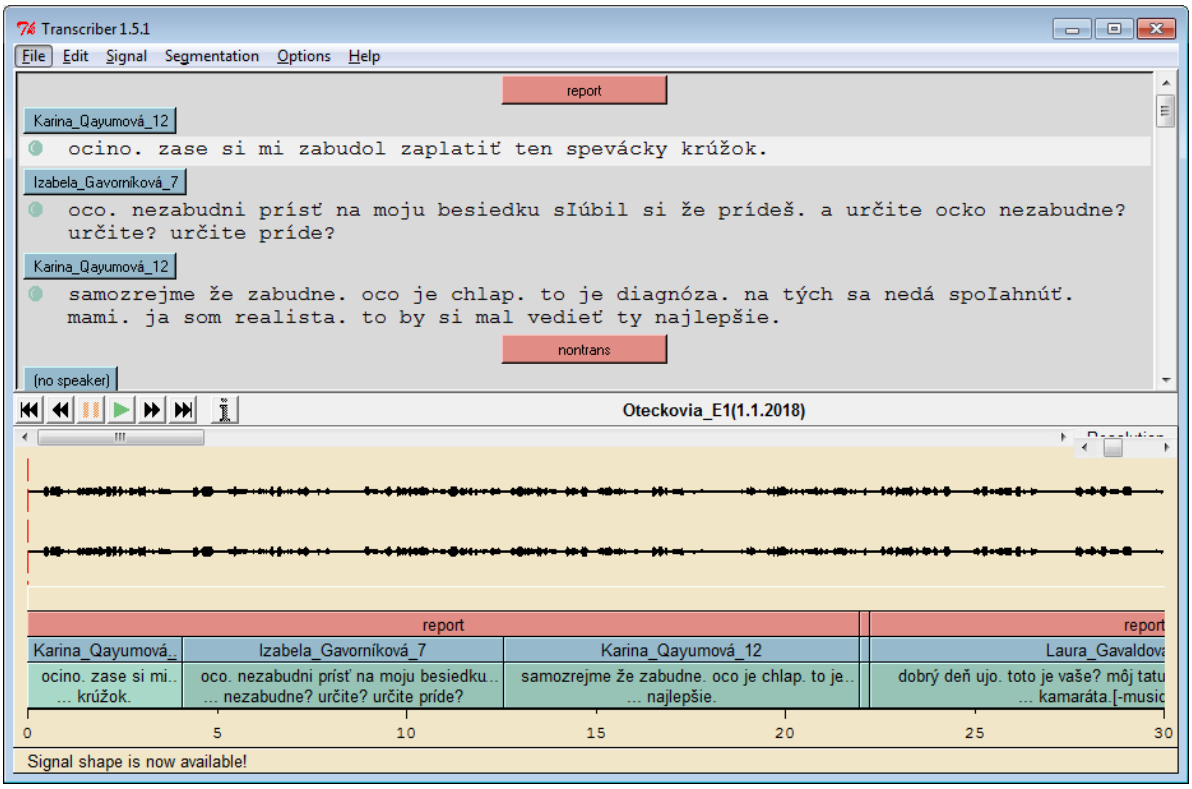

Figure 1. Transcription software used (Transcriber 1.5.1)

The gender, dialect, and the mother tongue (native or non-native speaker) were annotated for each speaker.

The mode field was set for speaker turns. We use the spontaneous option to indicate that the speech is spontaneous, unprepared speech or conversation. Mainly spontaneous speech was annotated for children. The planned speech is commonly used by studio moderators and sport news anchors. We follow the rules from standard broadcast news transcriptions [14] for the fidelity and the channel quality. Similarly, annotations mark the background noise and intermittent noise (see Figure 2.).

Annotations of the speaker turns follow the rule that one speaker turn should be no longer than 5 seconds. The capital letters at the beginning of the sentences were not used for easier named entity recognition.

The transcription process was made manually by the bachelor student and verified by his advisor. The plan is to extend the database following this proposed process in next year using more student annotators and expert verifications.

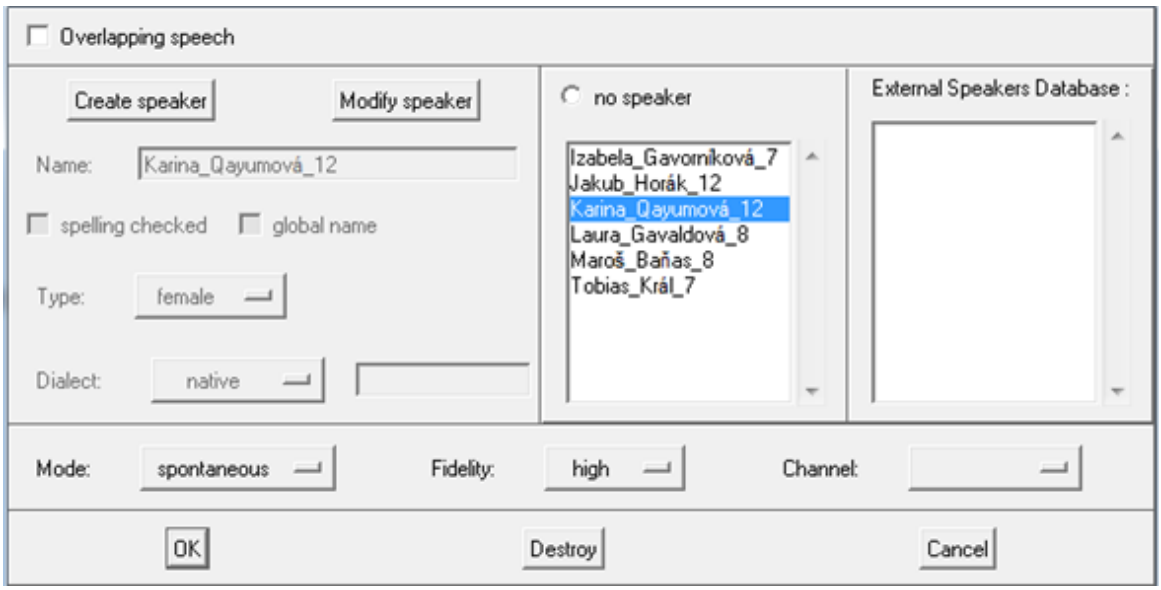

Figure 2. Transcriber window for speaker turn metadata.

4. Evaluation of the current subtitling system with children recordings

The current automatic subtitling system for Slovak TV broadcasters was developed thanks to many years of Slovak automatic speech recognition development of Technical University of Kosice and Slovak Academy of Sciences consortium. The previous system was based on Julius [10] mainly prepared for speech dictation into word processing editor. The next generation was built on Time Delay Deep Neural Network (TDNN) models based on Kaldi [11] for broadcast news transcription [17]. This version was also made online for public testing and evaluation on [18] as seen in Figure 3.

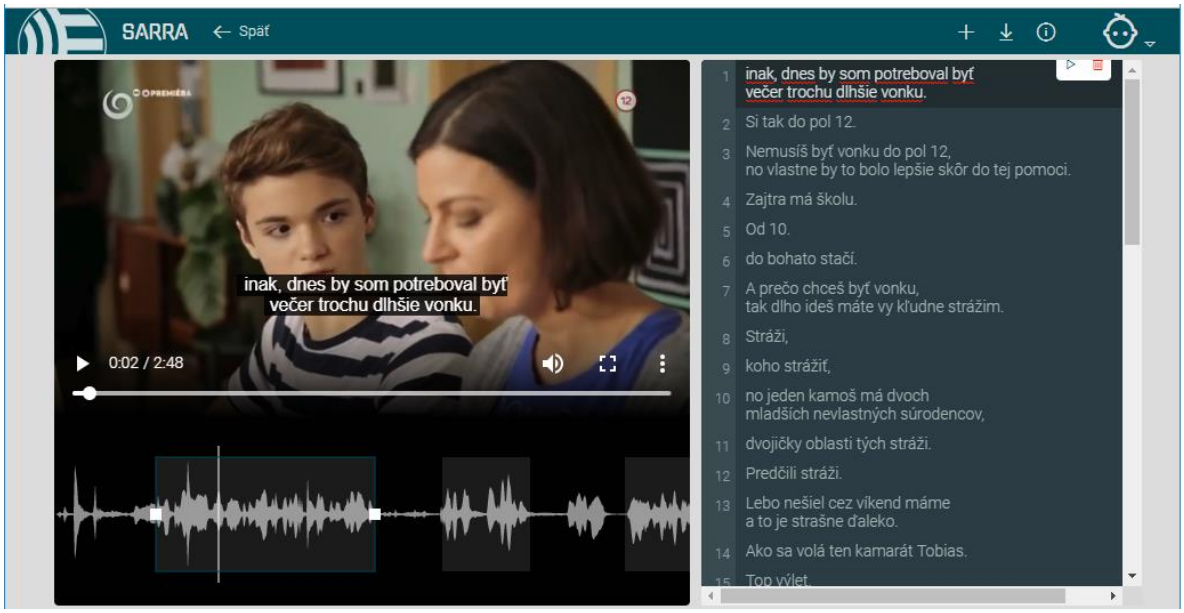

Figure 3. SARRA web user interface of automatically subtitled content

The SARRA system is built to work in multitasking and scaling environment, so the user's task could run on more instances of the recognition toolkit at once. The first part of the process is voice activity detection and speaker diarization for better segmentation of the large audio uploads. The smaller segmented parts of the audio could be scaled better.

The next part is the primary automatic speech recognition process built on models from about 600 hours of Slovak speech from broadcast news and TV discussions recordings. The acoustic model is using 40MFCC coefficients with online Cepstral mean normalization (CMN, in first training phase) and 100 dim i-vector. The language model was built from 1.89 billion token corpus with 500 thousand unique words vocabulary smoothed by the Witten-Bell algorithm [17].

The last part is the post-processing of the recognized text to convert it to the TV subtitle suitable form. The requirements were that the amount of text is limited on one subtitle caption and also the time for showing the caption should be long enough to read them by the viewers. Finally, it is expected to have the subtitles adapted to speaker changes where the diarization engine results are used.

The current engine could achieve 14.6% WER (Word Error Rate – the number of errors divided by the number of words in ground trough data) for broadcast news transcriptions where the variety of speakers and speech styles is wide [17]. For comparison, the dictation engine could achieve 3.93% WER for prepared Slovak dictation [10].

After the uploading and evaluation of the results from presented children Slovak speech database, we achieve only 47.81% WER, mainly because of 9.18% OOV (Out of vocabulary words) rate and very spontaneous speech segments.

5. Conclusions

The resources of children speech are scarce, even for major languages [7]. The presented database of children speech is the first one for the Slovak language and provides essential experiences about acoustic and mainly linguistic features of Slovak children speech. The development of adapted acoustic and language models for the Slovak automatic children speech recognition is in progress. There are several goals ahead, but mainly the extension of the presented dataset is planned for next year using more undergraduate students and expert verifications of the transcriptions. The goal is to present a special version of the SARRA models [17, 18] for children speech and evaluation by real users also for dictation and Human-Robot interaction purposes [19] based on the running international collaboration and projects.

Acknowledgments

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and finally by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 108-2911-I-027-501, 107-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067.

References

[1] Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A.: A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, ACM, p. 7, 2009.

[2] ISCA Special Interest Group: Child Computer Interaction (CHILD). [Online]. Available: https://www.isca-speech.org/iscaweb/index.php/sigs?id=129 [Accessed: July 30, 2019].

[3] Spoken Language Processing for Children's Speech, Interspeech 2019 Special session proposal. [Online]. Available: https://sites.google.com/view/wocci/home/interspeech-2019-special-session. [Accessed: July 30, 2019].

[4] Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., & Wong, M.: The PF_STAR children's speech corpus. In *Ninth European Conference on Speech Communication and Technology – INTERSPEECH 2005*. pp. 3761-3764, 2005.

[5] Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., Price, P., Anderson, E., Narayanan, S., & Alwan, A.: TBALL data collection: the making of a young children's speech corpus. In *Ninth European Conference on Speech Communication and Technology – INTERSPEECH 2005*. pp. 1581-1584, 2005.

[6] Xiangjun, D., & Yip, V.: A multimedia corpus of child Mandarin: The Tong corpus. *Journal of Chinese Linguistics*, 46(1), pp. 69-92, 2018.

[7] Wang, J., Ng, S. I., Tao, D., Ng, W. Y., & Lee, T.: A study on acoustic modeling for child speech based on multi-task learning. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP).* Taipei, IEEE, pp. 389-393, 2018.

[8] Watson, S., & Coy, A.: JAMLIT: A Corpus of Jamaican Standard English for Automatic Speech Recognition of Children's Speech. In *SLTU*. pp. 243-247, 2018.

[9] Pérez-Espinosa, H., Martínez-Miranda, J., Espinosa-Curiel, I., Rodríguez-Jacobo, J., Villaseñor-Pineda, L., & Avila-George, H.: IESC-Child: An Interactive Emotional Children’s Speech Corpus. *Computer Speech & Language*, *59*, pp. 55-74, 2020.

[10] Rusko, M. et al.: Advances in the Slovak Judicial Domain Dictation System. In: Vetulani

Z., Uszkoreit H., Kubis M. (eds) *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2013 - Revised selected papers*. Lecture Notes in Computer Science, vol 9561. Springer, Cham, pp. 55-67, 2016.

[11] Staš, J. et al.: Automatic subtitling system for transcription, archiving and indexing of Slovak audiovisual recordings. In *Proceedings of the 7th Language & Technology Conference, LTC 2015.* pp. 186-191, 2015.

[12] Lee, S., Potamianos, A., and Narayanan, S.: Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp.1455–1468, 1999.

[13] Shivakumar, P. G., Potamianos, A., Lee, S., and Narayanan, S.: Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In: Proc. *Workshop on Child, Computer and Interaction (WOCCI)*, pp. 15-19, 2014.

[14] Pleva, M. and Juhár, J.: TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation. In: *LREC*, Reykjavik, pp. 1709-1713, 2014.

[15] Fehér, M.: Automatic speech recognition for children. Bachelor thesis, Technical University of Kosice, p. 39, 2019.

[16] Transcriber - a tool for segmenting, labeling and transcribing speech. [Online]. Available: http://trans.sourceforge.net/en/presentation.php [Accessed: July 30, 2019].

[17] Lojka M., Viszlay P., Staš J., Hládek D., Juhár J.: Slovak Broadcast News Speech Recognition and Transcription System. In: Barolli L., Kryvinska N., Enokido T., Takizawa M. (eds) *Advances in Network-Based Information Systems*. NBiS 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 22. Springer, Cham, pp 385-394, 2019.

[18] SARRA - the automatic subtitling system for transcription, archiving, and indexing of Slovak audiovisual recordings. [Online]. Available: https://marhula.fei.tuke.sk/sarra/ [Accessed: July 30, 2019].

[19] Pleva, M., Juhar, J., Ondas, S., Hudson, C. R., Bethel, C. L., & Carruth, D. W.: Novice User Experiences with a Voice-Enabled Human-Robot Interaction Tool. In *2019 29th International Conference Radioelektronika*, Pardubice. IEEE, pp. 1-5, 2019.