



Live TV Subtitling through Respeaking

Aleš Pražák¹, Zdeněk Loose², Josef V. Psutka¹, Vlasta Radová¹, Josef Psutka¹, Jan Švec¹

¹Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic

²SpeechTech, s.r.o., Pilsen, Czech Republic

{aprazak, psutka-j, radova, psutka, honzas}@kky.zcu.cz, zdenek.loose@speechtech.cz

Abstract

In this paper, we describe our solution for live TV subtitling. The subtitling system uses the respeaking concept with respeakers closely tied with the automatic speech recognition system. The ASR is specially tailored to the live subtitling task by using respeaker-specific acoustic models and TV-show-dependent language models. The output stream of ASR could be online modified by keyboard shortcuts controlled by the respeaker. The whole subtitling service is used by Czech Television to provide high-quality subtitles of live shows for people with hearing impairments.

Index Terms: speech recognition, live TV show subtitling, respeaking

1. Introduction

Live TV subtitling is being increasingly demanded by the society of deaf and hard of hearing to make TV services accessible to people with hearing disabilities and the elderly. Live TV subtitles should convey the aural content of TV broadcasting in the original language to individuals who are deaf and hard of hearing to the same extent that the audio track conveys such content to individuals who can hear. Similar demand is enshrined in many international and national legislations. As opposed to offline subtitling with flawless and perfectly timed subtitles prepared in advance, live subtitling has many challenges. In addition to live subtitling coverage, which is approaching 100 percent in some countries such as the UK, Switzerland, and France, the focus is now placed on quality - latency and accuracy of live subtitles [1].

Since ASR technology is still unable to automatically transcribe an arbitrary TV audio with acceptable accuracy, the so-called respeaking (a technique in which a professional respeaker listens to the source audio and dictates it in a quiet environment to the well-tailored speech recognition system) has been established as the most widely adopted live subtitling technique. Live subtitling through respeaking was pioneered by British BBC in 2003, and it is now used worldwide in different implementations with different real-world parameters.

This show and tell paper describes our solution [2] for live subtitling through respeaking that differs from existing live subtitling platforms in the advanced technical level, resulting in higher live subtitling performance in terms of speed and accuracy with comparable costs. The specialized system is operated for several years by the SpeechTech company with close cooperation with the University of West Bohemia in Pilsen as a service for Czech Television (the public service broadcaster). The full list of subtitled shows is available online¹.

¹<https://www.zivetitulky.cz>

2. Respeaking concept

Respeaking is a complex technique to transcribe imperfect spoken messages and to provide intelligible and grammatically correct subtitles. A respeaker can repeat word-by-word if an original speech is clear and slow enough, so subtitle users are able to keep up. On the other hand, if multiple speakers are interrupting each other and speaking incoherently, verbatim transcription may not be comprehensible to hard-of-hearing viewers. In this case, a respeaker is expected to rephrase and/or condense the original speech by clear and grammatically correct sentences with the same meaning. An additional aim of the respeaker is to filter the information contained in sports commentaries (e.g., a possession of the puck in ice-hockey is visible) and to deliver to the viewer only important and interesting information in a form that does not bother nor distract the viewer.

It comes from the principle of the recognition system that the last few words of the recognized text change as the new acoustic signal is received and the best hypothesis based on acoustic and language model is recomputed. Since unchanging subtitles are preferred by deaf and hard of hearing, displaying of these last so-called “pending” words (four at maximum) is delayed during subtitle generation. However, these words can be displayed as highlighted to the respeaker, so he/she knows which words can be possibly corrected. The innovative correction of pending words is closely connected to the recognition system. In the case of misrecognition, the respeaker erases the pending words by a keyboard command and respeaks them fluently. The decoding process is terminated, the pending words are cut, and a new decoder hypothesis is established using the last dispatched words as context words for consecutive recognition. The described procedure requires a very low recognition latency; the words must be displayed to the respeaker within half a second after their utterance to allow the respeaker to check the text with minimum effort. If there is a potential OOV word that cannot be paraphrased, the idea is similar. The respeaker utters the word, and in case of misrecognition, the respeaker adds the word to the recognition system by typing it, erases misrecognized words, and respeaks added words once more.

The respeaker is also responsible for the insertion of correct punctuation marks. We do not use common spoken punctuation, instead, the respeaker uses his/her hands to press punctuation marks on the keyboard during inter-word pauses, so they can be processed by the recognition system that presents the punctuation marks directly in its result. As the language model considers punctuation marks as words, hypotheses scores are updated using language model probabilities of inserted words, so the recognition system can benefit from the extra non-speech information provided by the respeaker. The same approach is used for the insertion of speaker change markers using special tokens in the language model.

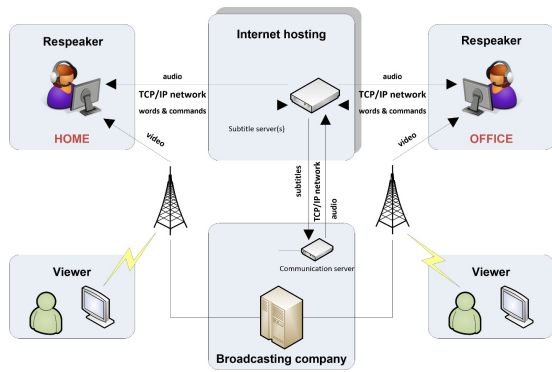


Figure 1: Architecture of remote live TV subtitling system employing respeakers.

3. Speech recognition

To achieve a recognition accuracy as high as possible, both the acoustic model and the language model of the recognition system must be tailored for individual respeakers and TV programs, respectively. The best way to train a speaker-specific acoustic model is to use only the target speaker's utterances obtained during live subtitling. To avoid manual transcriptions of such corpus, high-quality automatic transcriptions (by means of a recognition system) can be used for training. This leads to a cyclical problem that can be broken up during respeaker training when the speaker-independent acoustic model is used during the second training phase. After the first 100 training hours, the respeaker's utterances transcribed automatically during respeaker training are used for speaker-specific acoustic training; however, only words with very high confidence scores and highly credible neighboring words are used. This semi-supervised training process is repeated after each 100 training hours, gradually improving recognition accuracy. The ratio between automatically transcribed raw training data and selected real data slightly goes up with decreasing WER (commonly about 0.5). For acoustic modeling, we use HMMs with states modeled by the deep neural network trained in Kaldi using TDNN LF-MMI architecture on MFCC-based parameterized data.

To train specific language models for different TV program types (e.g. news, entertainment, chat shows, or sports), we use a large amount of training data from different sources. We have collected data from newspapers (480 million tokens), web news (535 million tokens), subtitles (225 million tokens), and transcriptions of some TV programs (210 million tokens) [3]. Since each sport has its specific terms and phrases that are commonly used during TV commentary of the sport, we manually transcribed TV commentaries of 40 different sports (e.g., baseball, golf, figure skating, or shooting) with 250K tokens per sport on average. The resulting trigram language models with mixed-case vocabularies incorporate over 1.4 million words for non-sports and 750K words for sports domains [4]. Even with such large vocabularies, a significant problem of live subtitling of sports TV programs are OOV words, especially the names of sportsmen and teams, which cannot be covered by any real training data [5]. However, participants in sports events are usually known in advance, so this problem leads to a class-based language model, where its 18 classes (trained on specially labeled sports data) should be filled before each live subtitling.

Since speech recognition decoders based on finite-state transducers are very difficult to modify on the fly, we use the phonetic prefix (lexical) tree structure of the decoding network

to allow online interventions to the whole decoding process. Lexical trees are highly efficient for languages with a high degree of inflection (such as the Czech language), where many word forms are derived from the same word stem. A time-synchronous Viterbi search on word-conditioned lexical tree copies is carried out. To enable recognition with a vocabulary containing more than one million words in real-time, the decoding process is highly parallelized by partitioning the vocabulary (and related lexical tree copies) to smaller units, their parallel decoding, and smart data synchronization. New words added to the vocabulary during the recognition by a respeaker are represented by an additional parallel unit that is simply integrated into the decoding process. Since trigram language model probabilities are factorized along the lexical trees on the fly, no pre-computing is required and new words can be recognized starting with the next time frame with full n-gram statistics. The whole system is operated on modern four-core laptop computers used by respeakers.

4. Conclusion

In this paper, we briefly described our approach to live subtitling through respeaking, characterized mainly by the very close connection between refined live subtitling platform and well-tailored speech recognition system. The system has many original features, such as instant corrections of misrecognitions, punctuation indication by keyboard, or new word addition during subtitling. Necessary elements also include speaker-specific acoustic models and domain class-based language models together with the full remote respeaking concept.

Even though respeaking, in our concept, is a highly demanding job, according to the real live subtitling sessions for Czech Television, one experienced respeaker can handle one to two hours of subtitling without a break. A respeaker training process is supported by our four-phase respeaker training system. We also provide some additional services related to the whole life cycle of live subtitles, such as a method for automatic live subtitle retiming (for TV program reruns) or a technique for live subtitle delay elimination.

5. Acknowledgement

This research was supported by the Technology Agency of the Czech Republic, project No. TN01000024.

6. References

- [1] P. Romero-Fresco, "Accessing communication: The quality of live subtitles in the UK," *Language and Communication*, vol. 49, pp. 56–69, 2016.
- [2] A. Pražák, Z. Loose, J. V. Psutka, V. Radová, and J. Psutka, "Live TV subtitling through respeaking with remote cutting-edge technology," *Multimedia Tools and Applications*, vol. 79, no. 1, pp. 1203–1220, 2020.
- [3] J. Švec, J. Lehečka, P. Ircing, L. Skorkovská, A. Pražák, J. Vavruška, P. Stanislav, and J. Hoidekr, "General framework for mining, processing and storing large amounts of electronic texts for language modeling purposes," *Language Resources and Evaluation*, vol. 48, no. 2, pp. 227–248, 2014.
- [4] J. Lehečka and A. Pražák, "Online LDA-Based Language Model Adaptation," in *Text, Speech, and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2018, pp. 334–341.
- [5] M. Hruží, A. Pražák, and M. Bušta, "Multimodal Name Recognition in Live TV Subtitling," in *Proc. Interspeech 2018*, 2018, pp. 3529–3532.