Novel Approach to Live Captioning Through Re-speaking:
Tailoring Speech Recognition to Re-speaker’s Needs

Ales Pražák\textsuperscript{1}, Zdeněk Loose\textsuperscript{2}, Jan Trmal\textsuperscript{2}, Josef V. Psutka\textsuperscript{2}, Josef Psutka\textsuperscript{1/2}

\textsuperscript{1}NTIS, University of West Bohemia, Plzeň, Czech Republic
\textsuperscript{2}Department of Cybernetics, University of West Bohemia, Plzeň, Czech Republic

aprazak@kky.zcu.cz, \{zloose, jtrmal, psutka_j, psutka\}@kky.zcu.cz

Abstract

A novel approach to the live captioning through re-speaking is introduced in this paper. We describe our concept of re-speaking using only one re-speaker with enhanced re-speaking tasks fully integrated to the recognition system and captioning software. New techniques for instant correction of recognition output, punctuation mark introduction or new word addition are presented. Our real-time recognition system of the Czech language with a vocabulary containing more than one million words is described and an architecture of captioning system that we operate is illustrated. Last part of the paper is dedicated to the re-speaker training methodology and a three-level evaluation method of final re-speaker’s skills is proposed.

Index Terms: live captioning, speech recognition, re-speaking

1. Introduction

Television companies all around the world are pushed by the law and the society of deaf and hard of hearing to provide closed captions to 100\% of their broadcasts. To satisfy these demands, television companies turn to private companies (will be referred to as caption companies) which are familiar with offline caption preparation through translators and transcribers. Television companies require 24/7 captioning including live TV programs in the same way as for offline programs with captions prepared in advance - verbatim transcription with precisely-timed captions. According to last trends (re-speaking as a cost-effective alternative over keyboards and stenotyping), a caption company buys a speech recognition software as a supplement to its existing captioning platform and employs its translators as re-speakers and transcribers as correctors.

It surely is the way how to do live captioning. An example is the British BBC that starts as a pioneer with live captioning in 2003 \cite{1, 2} and now its caption company Red Bee Media employs about 200 re-speakers/correctors for 24/7 live captioning of 7 BBC channels. However, there are obvious drawbacks of such an approach, since a speech recognition software acts only as a black box for these companies. For example IBM’s ViaVoice serves only for alternative (speech) input to the SysMedia’s WinCAPS, basically offline captioning platform. Consequently, a re-speaker serves merely as a human machine that re-speaks original speech word-by-word with no feedback or interaction with restricted speech recognition software. In other words, such dictating systems are not primarily designed to be used for live captioning. To overcome still imperfect recognition performance and provide intelligible captions with punctuation and speaker coloring, one or more correctors must be used. This implies expensive and quite unwieldy solution, because the speech recognition technology is not fully applied.

Unlike former captioning methods through keyboards (Velotype, stenotype, etc.), speech recognition provides higher incessant transcription speed, so verbatim transcription can be accomplished. On the other hand, there are many people among the target audience of closed captions (for example elderly people), who have limited reading speed and captions with more than 180 words per minute are frustrating for them \cite{3}. In addition, some of the deaf people (mainly deaf from birth) may have limited vocabulary and they poorly comprehend the transcription of grammatically incorrect speech.

Since we have been developing speech recognition systems in conjunction with live captioning applications for many years, we have proposed a novel approach to the re-speaking that combines editing of the original speech by the re-speaker (in fact, intralinguistic simultaneous translation including reduction and rephrasing) and full interaction of the re-speaker with the recognition system and captioning software (including instant correction of misrecognitions, punctuation indication, new word addition, etc.). Thus this approach combines re-speaker and corrector tasks into one highly-skilled profession that is cheaper than two co-workers but requires quite high qualification.

During our cooperation with the Czech Televison, the public service broadcaster in the Czech Republic, we have developed and implemented a system for automatic live captioning of the meetings of the Parliament of the Czech Republic directly from the real acoustic track \cite{4}. We operate such a system over four years with almost 1,000 captioned hours. The recognition accuracy reaches 90\% in average (depending on the topic discussed). Since re-speaking brings up new practical issues, we have dealt with these issues at the level of speech recognition to employ all its potential in the real application. This paper describes our complete solution for live captioning through re-speaking from the recognition system modifications to the distributed captioning system architecture and re-speaker training methodology.

2. Speech recognition

Since the speech recognition system is closely connected to non-speech tasks of the re-speaker, speech recognition details should be described first.

We employ in-house LVCSR system optimized for low-latency real-time operation with very large vocabularies on multi-core systems. While a speech decoder based on finite-state transducer does not support simple vocabulary or language model modifications just during recognition, we use a lexical (phonetic prefix) tree structure of a recognition network. 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. To enable recognition with a vocabulary containing more than one million words in real-time, we use Hidden Markov Models with static triphone in-word context and dynamically generated triphone cross-word context. A Viterbi search on word-conditioned lexical tree copies is used. The decoding process is highly parallelized by partitioning the vocabulary (and related lexical trees) to the smaller units, their parallel decoding and smart data synchronization. Words added to the vocabulary during the recognition by a re-speaker are represented by additional parallel unit which is simply integrated to the decoding process. Since n-gram 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.

To achieve the best recognition performance, the language model of the recognition system should be specific for each type of live program (e.g. political debates, TV shows, sport programs). Therefore we need a huge amount of language model training data from different sources of different nature. We have collected large corpus containing the data from newspapers (520 million tokens), web news (350 million tokens), subtitles (200 million tokens) and transcriptions of some TV programs (175 million tokens) [5]. These data are automatically processed using topic detection and updated daily, so we have topic specific language models with up-to-date vocabularies.

Now we use trigram back-off language models (using Kneser-Ney discounting) with mixed-case vocabularies containing more than one million words; however, even such models cannot cover all the words of the language with a high degree of inflection (with a large number of prefixes and suffixes). To cover common OOV words, we have prepared special large lists of named entities. These lists include for example Czech surnames, Czech villages and streets, world countries and cities; over 200,000 items in all grammatical cases and other word forms (more than ten times more unique words) with pronunciation variants. These words are not present in the active vocabulary, but they can be added to the vocabulary by the re-speaker during recognition with full trigram statistics based on classes trained for each list and word form.

Since a re-speaker’s voice is being recognized instead of real acoustic track of a TV program, speaker-dependent acoustic model should be trained. We train such a model for each re-speaker in two steps. In the first step, we use a gender-dependent acoustic model (trained on 150 hours from 500 speakers), that is automatically adapted to the re-speaker’s voice characteristics online during recognition. An unsupervised incremental fMLLR adaptation is carried out in the background based on recognized words and their confidence scores [6]. Only words with very high confidence scores and highly credible neighboring words are used for adaptation. This approach can improve recognition performance until enough speech data (30 hours at minimum) is gathered. In the second step, a speaker-dependent acoustic model is trained unsupervised based on gathered data, their recognized transcriptions and word confidence scores. We use triphone acoustic models with variable number of components with PLP parameterization (19 filters, 12 cepstral coefficients with both delta and delta-delta sub-features, 100 frames per second) on 22 kHz acoustic signal. Acoustic model computation is carried out on graphic processor unit (GPU) [7], so due to high efficient decoder parallelization we can recognize with a vocabulary containing over one million words in real-time on modern four-core laptop computers.

3. Re-speaking

Re-speaking in our concept is a complex technique to transcribe imperfect spoken message and provide intelligible and grammatically correct captions. A re-speaker can repeat word-by-word if an original speech is clear and slow enough, so viewers are able to keep up. On the other hand, if more speakers are interrupting each other and speaking incoherently, verbatim transcription may not be comprehensible by hard-of-hearing viewers. In this case a re-speaker is expected to rephrase the original speech by clear and grammatically correct sentences with the same semantic meaning. A reduction of the original message includes omission of discourse markers, intensifiers, asides or repetitions of already stated information. On the other hand, one caption usually does not cover the entire sentence, let alone the whole speaker idea. For example a simple speaker answer “Yes” should be amended by a re-speaker to the full phrase “Yes, I was ...” to remind the original idea to a viewer, if necessary.

It comes from the principle of the recognition system that last few words of the recognized text change as new acoustic signal is received and the best hypothesis based on acoustic and language model is recomputed. Since static closed captions are preferred by deaf and hard of hearing, these last so-called “pending” words (four at maximum) are ignored during the caption generation. However, these words can be displayed to the re-speaker highlighted, so he/she knows which words can be possibly corrected. The correction of pending words is closely connected to the recognition system. In case of misrecognition, re-speaker erases the pending words by a keyboard command and re-speaks them fluently. The decoding process is terminated, the pending words are cut and a new decoder hypothesis is established using last dispatched words as context words for consecutive recognition. If there is a potential OOV word which cannot be paraphrased, the idea is the same. Re-speaker utters the word and in case of misrecognition the re-speaker adds the word to the recognition system by typing it, erases misrecognized words and re-speaks added word once more. Moreover, if an added word was found in our large lists of named entities, the word and all its word forms with pronunciation variants are added to the vocabulary, so from now on the named entity in all its word forms will be recognized correctly.

A time lag of the final caption is dictated by the length of the caption and the pending word issue. It can be reduced if the re-speaker manually dispatches the pending words. This is very important when the re-speaker is about to be silent for a longer time (because of a TV jingle or the fact that he/she is listening ahead) and the pending words remain absent. Based on keyboard command, the recognition system evaluates and dispatches current best recognition hypothesis including pending words and starts a new recognition with the context carried over. This can significantly reduce the time lag between the words uttered in the original speech and corresponding words in the captions.

The only way how to segment continuous recognition output in existing live captioning systems, is to use a post-editing by a corrector, or to use an oral punctuation by a re-speaker. Obviously, an oral punctuation reduces the transcription speed of re-speaking. In our approach, a re-speaker uses his/her hands to press punctuation marks on keyboard during inter-word pauses, so they can be processed by the recognition system that presents the punctuation marks directly in its result. An artificial insertion of a word to the recognition system is based on pruning at the level of lexical trees. Since a word can be inserted only during inter-word pause, all active word hypotheses at the lex-
ical trees are pruned, while hypotheses in non-speech models survive supplemented by inserted word. 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 extra information provided by the re-speaker. The same approach is used for insertion of speaker change markers using special tokens in the language model.

Even though re-speaking in our concept is highly demanding job, according to the real live captioning sessions for the Czech Television, one experienced re-speaker can handle one to two hours of captioning without a break.

4. Captioning system

A time lag of the final caption given by its processing is about 3-5 seconds, moreover the transmission of the caption to the viewer takes another 3 seconds. To avoid the time lag completely, the broadcaster would have to hold up the broadcasting up to 10 seconds to provide precisely-timed captions. Since television companies are not ready to do this, we receive the real acoustic track of the TV program ahead of its transmission, so we can eliminate at least above mentioned 3 seconds of the time lag.

Because of the strict electronic security policies at the Czech Television (CTV) in Prague, the captioning system architecture was designed as highly distributed with the centre at the University of West Bohemia (UWB) in Pilsen (see Figure 1). The interconnection between CTV and UWB is done by a point-to-point connection over the ISDN network. To be able to carry both audio signal to UWB and generated captions to CTV, the ISDN is used only as a full-duplex data carrier with bandwidth of 128 kbit/s. Specialized terminal adapters commonly used in CTV with transparent low-latency compression are used. The captioning server at UWB distributes audio signal by VoIP service to the re-speakers to arbitrary locations with internet connection. Since a visual component of a TV program is not delivered, common DVB signal is displayed to the re-speaker. The synchronization delay of 3 seconds between audio and video is not crucial, because the visual component is intended only for re-speaker’s overview about the situation.

The whole interactive recognition process is carried out on a laptop computer. Since boxed pop-on closed captions are demanded, formatting of captions is done automatically at the captioning server. To trade-off between the caption time lag and its length, one row captions are preferred, but other requirements related to the readability of the final captions are considered too. Based on speaker change markers provided by the re-speaker, an automatic speaker coloring of final captions is performed.

5. Re-speaker training

Since our re-speaking philosophy assumes comprehensive re-speaker’s skills, a complex training process is required. Such a training should be carried out using the real captioning software, so that the re-speaker becomes accustomed to it; on the other hand, a re-speaker candidate should be prepared gradually from the basics to the most demanding tasks. Therefore we have developed a special system for re-speakers that provides gradual training process under surveillance of a skilled supervisor, so the training can be shorter and cheaper. Four phases are intended for training, while the fifth phase serves for the real captioning process.

The first training phase is intended to train candidate’s skill to listen and speak simultaneously. The candidate opens prepared video file and practices speaking while playing any part of the video. The aim is not to re-speak word-by-word, but become accustomed to speaking meaningfully while listening to and perceiving the original acoustic track. This phase does not employ recognition system, but all utterances are recorded for later playback by the candidate and inspection by the supervisor. The first phase helps to sort out re-speaker candidates which do not have basic prerequisites to do the job.

The second phase of the training system assists in optimizing the candidate’s speech to the recognition system demands, so this phase integrates a LVCSR system and displays its output to the candidate. The main objective of the candidate is to re-speak original speech word-by-word so that the recognition accuracy is as high as possible. The candidate just mechanically re-speaks what he/she listens to, so he/she can focus on altering the utterance (mainly the pronunciation) and its influence on the recognition results. Since a transcription of demanded candidate’s utterance is known in advance, the misrecognitions in the recognized text can be highlighted based on Levenshtein alignment. In addition, overall recognition accuracy is displayed, so the candidate’s training progress can be monitored.

In the third phase of the training system, a trainee learns rephrasing and condensing whenever it is required. A candidate has to check instantly recognized text and perform potential corrections of pending words. Other keyboard commands should be used to dispatch pending words, to add new words to the vocabulary and to mark speaker changes too. The fourth phase simulates the real captioning system with all the features needed during the real caption generation including punctuation mark insertion.

The overall duration of a re-speaker training is very individual, but according to our expertise, an average time of intensive training plan is from 2 to 3 months (75 training hours at minimum). For comparison, a broadcast stenography requires two or three years of training and experience, resulting into a chronic shortage of skilled stenographers.

6. Re-speaker evaluation

Due to rephrasing and condensing of the original message by a re-speaker, a speech recognition output cannot be directly compared with verbatim transcription of original speech in a TV program. To assess final re-speaker’s skills, we have proposed a three-level evaluation method. A re-speaker re-speaks a real

Figure 1: Captioning system architecture.
TV program in the fifth phase of the captioning system and final captions (thus including corrections made by the re-speaker) are then evaluated in the following three levels.

The first level is represented by standard recognition accuracy, except the fact that misrecognitions of the slips of the tongue which result back in the intended words are considered as correct words. This may have considerable impact on the recognition accuracy, because a language model of the recognition system often rectifies the re-speaker’s pronunciation mistakes. The punctuation marks are not included in the evaluation.

In the second level, the syntactic correctness is evaluated. The reference transcription from the first level is further revised to contain only grammatically and lexically correct sentences (using minimum number of corrections). Only the word order in each sentence apart from the sentence context is usually revised. The resulted reference reflects grammatically correct sentences with the meaning of the original re-speaker utterance. The recognition accuracy based on the revised reference transcription is then evaluated.

Finally, the semantic meaning is evaluated in the third level. The original dialogues of a TV program are divided into so-called “turns”. One turn usually comprises one speaker utterance. Now each turn is rated upon the corresponding part of the re-speaker utterance by the following scores:

3: the re-speaker said correctly all the essentials (including the case when nothing was said because there was nothing essential)
2: the re-speaker said all the essentials unclearly or ambiguously (even if the original speaker did the same)
1: the re-speaker missed something essential from the original message
0: the essentials were totally missed or something wrong was added

The rating of the third level is then computed by the simple formula

\[ 3^{rd\ level\ rating} = \frac{\sum \text{turn score}}{3 \cdot \text{turn number}} \cdot 2^{nd\ level\ rating} \]

The ratings of the lower levels are iteratively used in the higher levels, so the rating in the third level represents overall error rate including recognition, syntactic and semantic errors, in other words a percentual ability to express original ideas in the text form (by means of the recognition system). Naturally, the evaluation in the third level is largely subjective.

Since a behavior of original speakers during a TV program can be very uneven, we have partitioned the ratings for two of our five skilled re-speakers to five parts (see Figure 2). There is a visible difference in the ratings of the first and higher levels between re-speakers EK and DT. This is caused by different approaches to the re-speaking. DT re-speaks in almost verbatim style (5 082 words, 128 corrections), it means nearly word-by-word omitting only slips of the tongue and some semantic fillings. This approach copies a sentence structure and fragmentation of the original speech. Hence the drop in the second and third level rating. On the other hand, EK uses listening ahead (4 572 words, 63 corrections), so she much more rephrases and creates her own grammatically correct sentences with the meaning of the original speech. However, such an approach can be inconvenient in case of fast speaker changes (from 30 to 40 minutes, see Figure 2).

7. Conclusions

We have presented our novel approach to the live captioning through re-speaking. Re-speaking in our concept is a complex technique to transcribe imperfect spoken message and provide intelligible and grammatically correct captions. Extended re-speaker tasks are fully integrated to the recognition system and captioning software, so one experienced re-speaker can reach caption quality of two co-workers (re-speaker and corrector) in existing live captioning systems. We have proved our concept on over 150 hours broadcasted by the Czech Television and live captioned by our five re-speakers, who got through the whole proposed training process.

8. Acknowledgements

This paper was supported by the Technology Agency of the Czech Republic, project No. TA01011264 and by the European Regional Development Fund (ERDF), project “New Technologies for Information Society” (NTIS), European Centre of Excellence, ED1.1.00/02.0090. This support is gratefully acknowledged.

9. References