Towards Live Subtitling of TV Ice-hockey Commentary

Aleš Pražák[1], Josef V. Psutka[2], Josef Psutka[2] and Zdeněk Loose[2]
[1]SpeechTech s.r.o., Plzeň, Czech Republic
[2]Department of Cybernetics, University of West Bohemia, Plzeň, Czech Republic

Keywords: Speech Recognition, Live TV Subtitling, Re-speaking, Sport.

Abstract: This paper deals with live subtitling of TV ice-hockey commentaries using automatic speech recognition technology. Two methods are presented - a direct transcription of a TV program and a re-speaking approach. Practical issues emerging from the real subtitling system are introduced and their solutions are proposed. Acoustic and language modelling is described as well as modifications of existing live subtitling application. Both methods were evaluated during simulated subtitling and their results are discussed.

1 INTRODUCTION

Recently, the public service televisions all around the world are pushed by the law and the society of deaf and hard-of-hearing to provide subtitles to all their broadcasts. For shows that are broadcasted offline is not a problem to provide them with offline subtitles through human translators and transcribers. But the subtitles for directly broadcasted programs are still a big challenge.

The large vocabulary continuous speech recognition (LVCSR) systems are demanded as a solution. The advantage of using automatic speech recognition (ASR) is a high transcription speed in comparison with human typing speed and a cheaper operation. On the other hand, the ASR system has not achieved such accuracy as human transcribers yet.

In the task of live TV subtitling, an ASR system can be used in two principal ways. The first one employs the speech recognition for direct transcription of an audio track of a TV program. This requires clear speech with calm acoustic background and non-overlapping speakers. Since only a few TV shows are suitable for direct transcription, a more universal strategy employs so-called "re-speaker" - a specialized speaker, who re-speaks the original dialogues for the speech recognition system (Evans, 2003).

During our cooperation with the Czech Television, the public service broadcaster in the Czech Republic, we have developed and implemented a system for automatic live subtitling of the meetings of the Parliament of the Czech Republic directly from the real audio track (Trmal et al., 2010). We operate such system over four years with almost 1,000 subtitled hours. The recognition accuracy reaches 90 % in average (depending on the topic discussed).

In recent years, we developed a distributed system for the live subtitling through re-speaking. A unique four phase re-speaker training system was introduced in (Pražák et al., 2011). We use such a system in cooperation with the Czech Television for live subtitling of political debates and similar TV shows.

Since subtitling of live sport programs is demanded by deaf and hard-of-hearing, we have faced new challenges and tasks to be solved. The most watched sports in the Czech Republic are ice-hockey and football. Initially, we tried to directly recognize a live TV commentary of an ice-hockey match. We had to prepare a specialized class-based language models covering names of players, teams or sport places and train dedicated acoustic model for the speech of the TV ice-hockey commentator. We did not expect too much from the direct recognition, while the re-speaker approach was more promising. To prove our assumption, we modified existing re-speaking application to make subtitling of live ice-hockey matches through re-speaker possible.

Pražák A., Psutka J., Psutka J. and Loose Z..
Towards Live Subtitling of TV Ice-hockey Commentary.
DOI: 10.5220/0004606301510155
In *Proceedings of the 10th International Conference on Signal Processing and Multimedia Applications and 10th International Conference on Wireless Information Networks and Systems* (SIGMAP-2013), pages 151-155
ISBN: 978-989-8565-74-7

2 LANGUAGE MODEL TUNING

In fact, each sport is unique with specific expressions, phrases and a manner of speech of the TV commentator. That is why a language model should be based on many transcriptions of the TV commentary of the given sport.

We did manual transcriptions of 90 ice-hockey matches, both international and Czech league matches. These transcriptions contain the names of players and teams involved in the match, but also other names that the commentators talked about. To make a general language model suitable for all ice-hockey matches, we would need to add all names of ice-hockey players in the world. This would swell the vocabulary of the recognition system and slow the system down, moreover the accuracy of transcription would drop. The only way is to prepare a language model specifically for each match by adding only names of players of two competing teams. A class-based language model takes the role in this task.

During manual transcriptions of TV ice-hockey commentaries, some words were labelled with the tags representing several semantic classes. The first class represents the names of players that take part in the match. The second class is used for the names of competing teams or countries and the next class for the designations of sport places (stadium, arena etc.). The names which do not relate to the transcribed ice-hockey match were not labelled (for example legendary players like "Jagr"), because they are more or less independent of the match. Since the Czech language is highly inflectional, further 27 classes were used for the names in other grammatical cases and their possessive forms. Finally, taking into account the above mentioned tags instead of the individual names, two class-based trigram language models were trained - one for in-game commentary and one for studio interviews.

The manual transcriptions of 90 commentaries contain 750k tokens with 25k unique words. These data cannot cover commentary of forthcoming ice-hockey matches. To make the vocabulary and language model more robust, other data from newspapers (175M tokens) and TV news transcriptions (9M tokens) were used. Only data with automatically detected topic of sport (Skorkovská et al., 2011) were used and mixed with ice-hockey commentaries based on perplexity of the test data. For in-game language model, the weights were 0.65 for ice-hockey commentaries, 0.30 for newspaper data and 0.05 for TV news transcriptions, while for studio interviews the weights were 0.20, 0.65 and 0.15, respectively. The final vocabulary of the recognition system contains 455k unique words (517k baseforms).

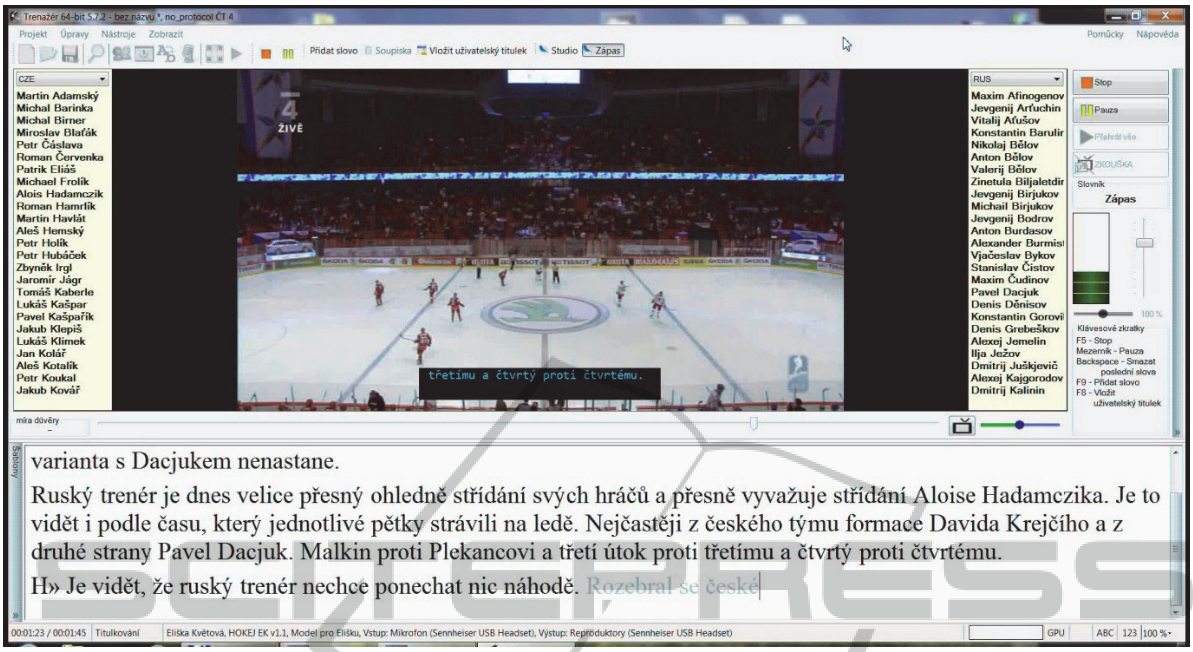
Finally, before recognition of each ice-hockey match, the language model classes have to be filled with the actual words. The names of players of two competing teams (line-ups) are acquired and automatically declined into all possible word forms. Since the player can be referred to by his full name or surname only, both representations are generated. Other language model classes are filled by the names of teams and designations of sport places corresponding to the given ice-hockey match.

3 DIRECT RECOGNITION

The acoustic data for direct subtitling (subtitling from the original audio track) was collected over several years especially from the Ice-hockey World Championships as well as from the Winter Olympic Games and the Czech Ice-hockey League matches. All these matches were broadcasted by the Czech Television. Sixty nine matches were transcribed for the acoustic modelling purposes. All these matches were manually annotated and carefully revised (using annotation software Transcriber). Total amount of data was more than 100 hours of speech.

The digitalization of an analogue signal was provided at 44.1 kHz sample rate, 16-bit resolution. The front-end processor was based on the PLP parameterization (Hermansky, 1990) with 27 band pass filters and 16 cepstral coefficients with both delta and delta-delta sub-features. Therefore one feature vector contains 48 coefficients. The feature vectors were calculated each 10ms. Many noise reduction techniques were tested to compensate for very intense background noise. The J-RASTA (Koehler et al., 1994) seems to be the best noise-reduction technique in our case (see details in Psutka et al., 2003).

The individual basic speech unit in all our experiments was represented by a three-state HMM with a continuous output probability density function assigned to each state. As the number of possible Czech triphones is too large, phonetic decision trees were used to tie states of the triphones. Several experiments were performed to determine the best recognition results depending on the number of clustered states and also on the number of mixtures per state. The best recognition results were achieved for 16 mixtures of multivariate Gaussians for each of 7700 states (see Psutka, 2007 for methodology).



Martin Adamský
Michal Barinka
Michal Birner
Miroslav Blaťák
Petr Čáslava
Roman Červenka
Patrik Eliáš
Michael Frolík
Alois Hadamczik
Roman Hamrlík
Martin Havlát
Aleš Hemský
Petr Holík
Petr Hubáček
Zbyněk Irgl
Jaromír Jágr
Tomáš Kaberle
Lukáš Kašpar
Pavel Kašpařík
Jakub Klepiš
Lukáš Klimek
Jan Kolář
Aleš Kotalík
Petr Koukal
Jakub Kovář

Maxim Afinogenov
Jevgenij Arťuchin
Vitalij Aťušov
Konstantin Barulir
Nikolaj Bělov
Anton Bělov
Valerij Bělov
Zinetula Biljaletdir
Jevgenij Birjukov
Michail Birjukov
Jevgenij Bodrov
Anton Burdasov
Alexander Burmist
Vjačeslav Bykov
Stanislav Čistov
Maxim Čudinov
Pavel Dacjuk
Denis Děnisov
Konstantin Gorovil
Denis Grebeškov
Alexej Jemelin
Ilja Ježov
Dmitrij Juškjevič
Alexej Kajgorodov
Dmitrij Kalinin

Klávesové zkratky
F5 - Stop
Mezerník - Pauza
Backspace - Smazat poslední slova
F9 - Přidat slovo
F8 - Vložit uživatelský titulek

třetímu a čtvrtý proti čtvrtému.

varianta s Dacjukem nenastane.

Ruský trenér je dnes velice přesný ohledně střídání svých hráčů a přesně vyvažuje střídání Aloise Hadamczika. Je to vidět i podle času, který jednotlivé pětky strávili na ledě. Nejčastěji z českého týmu formace Davida Krejčího a z druhé strany Pavel Dacjuk. Malkin proti Plekancovi a třetí útok proti třetímu a čtvrtý proti čtvrtému.

H» Je vidět, že ruský trenér nechce ponechat nic náhodě. Rozebral se české

00:01:23 / 00:01:45 | Titulkování | Eliška Květová, HOKEJ EK v1.1, Model pro Elišku, Vstup: Mikrofon (Sennheiser USB Headset), Výstup: Reproduktory (Sennheiser USB Headset)

Figure 1: Re-speaking application for ice-hockey matches.

A TV ice-hockey match is usually commented by the main commentator together with the co-commentator. The majority of the annotated matches was accompanied by the commentary by Robert Záruba, one of the most experienced Czech ice-hockey commentators. Because of a limited and relatively small set of speakers, the speaker-adaptive models (SAT) were trained for each speaker via fMLLR, to improve the recognition accuracy (Zajíc et al., 2009).

4 RE-SPEAKING

Unlike the direct recognition of the original audio track of a TV program, a re-speaker can reach higher transcription accuracy and produce easily readable and intelligible subtitles.

The main objective of the re-speaker is to listen to the original speakers and re-speak their talks. Unlike former subtitling methods through keyboards (Velotype, stenotype, etc.), speech recognition provides higher (and more stable) transcription speed, so verbatim transcription can be accomplished. On the other hand, there are many people among the target audience of subtitles (for example elderly people), who have limited reading speed and subtitles with more than 180 words per minute are frustrating for them (Romero-Fresco, 2009). Similar problems are raised in case of incoherent speech or overlapping speakers. In these cases, the re-speaker is expected to simplify and rephrase the original speech by clear and grammatically correct sentences with the same semantic meaning, so viewers are able to keep up.

Although speaker-specific acoustic model is used and the recognition accuracy reaches up to 98 %, crucial misrecognitions may appear and should be corrected by the re-speaker. Since static closed captions are preferred by deaf and hard-of-hearing, latest recognized so-called ”pending” words (four at maximum) are ignored during the subtitle generation. In case of misrecognition, re-speaker erases the pending words by a keyboard command and re-speaks them fluently. Moreover in case of out-of-vocabulary word, the re-speaker simply adds new word to the recognition system by typing it. Other keyboard commands are used for punctuation marks and speaker change indication to produce easily readable subtitles with punctuation and speaker colouring.

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

To make the re-speaking of ice-hockey matches possible, we had to modify our re-speaking application. First, the language model classes have to be filled instantly by the re-speaker just before the subtitling to avoid time-consuming supervised building of the language model and its transmission to the re-speaker. Before each match starts, the re-speaker chooses predefined team line-ups, team names and sport places which are dynamically added to the vocabulary and class-based language model of the recognition system. Team line-ups are than displayed beside the video for the case, when the re-speaker is uncertain about the names (see Figure 1).

Since only one language model is employed during subtitling of political debates, we have implemented an instant switching of different language models for in-game commentary and studio interviews. By a special keyboard command, the re-speaker chooses the appropriate language model after switching from the hall to the studio and vice versa. This approach reduces the recognition errors, because in-game commentary and studio interviews are pretty different from the language modelling point of view.

Last but not least, re-speaker behaviour has changed with respect to the lag of the final subtitles (see below). Some in-game situations may be irrelevant by the time the subtitle is displayed to the viewer. Furthermore, redundant subtitles during the play (e.g. a possession of the puck is visible) excessively occupy the viewer. Based on the experience of deaf and hard-of-hearing viewers, our re-speaker filters such superfluous information and focuses on rich commentary of interesting situations not covered by the video content.

5 SUBTITLING SYSTEM

Because of the strict electronic security policies at the Czech Television (CTV) in Prague, the subtitling system architecture was designed as highly distributed with the centre at the University of West Bohemia (UWB) in Pilsen (see Figure 2). 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 subtitles to CTV, the ISDN is used only as a full-duplex data carrier with a bandwidth of 128 kbit/s. Specialized terminal adapters commonly used in CTV with transparent low-latency compression are used. The subtitling server at UWB distributes audio signal by VoIP service to the re-speakers to arbitrary locations with reliable internet connection. Since a visual component of a TV program is not delivered, common DVB signal is displayed to the re-speaker; however it is intended only for re-speaker's overview about the situation. The re-speakers employ high-performance laptop computers to do their job. Based on words and commands sent back to UWB, subtitling server generates subtitles coloured according to the speaker change markers and supplies CTV broadcasting station.

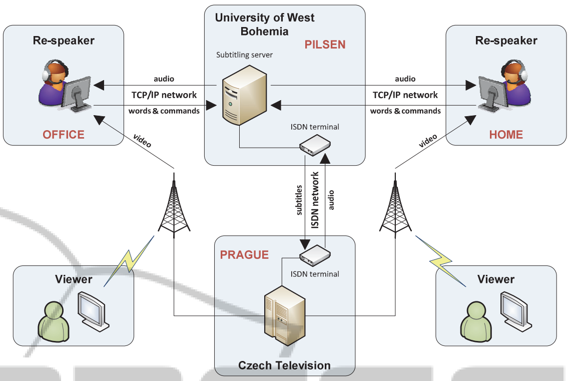


Figure 2: Subtitling system architecture.

Since static pop-on subtitles (closed captions) are preferred by deaf and hard-of-hearing in the Czech Republic, a time lag of the final subtitle is dictated by the length of the subtitle (all words must be uttered before the subtitle can be generated). Moreover, the processing and transmission of the subtitle to the viewer takes another 5 seconds. To avoid the time lag completely, the broadcaster would have to hold up the broadcasting up to 10 seconds to provide precisely-timed subtitles. Since television companies are not ready to do this, we receive at least the real audio track of the TV program about 3 seconds ahead of its transmission, so the final time lag on the viewer's side is about 5 seconds.

6 EXPERIMENTAL RESULTS

To test the proposed methods, one match from the last Ice-hockey World Championship was re-spoken by a skilled re-speaker and his speech was manually transcribed as a reference for the re-speaking approach. The original audio track of ice-hockey match was also transcribed to serve as reference for the direct recognition. This match was used neither for acoustic nor language modelling. The test match is 189 minutes long and it contains both the interviews in the studio before and after the match and during the breaks as well as in-game commentary. The acoustic model described above was used for the direct recognition, while an acoustic model designed for the re-speaker was employed during re-speaking. In both cases, the same trigram class-based language model was applied.

Word error rate (WER) of recognition and final subtitles (including corrections performed by re-speaker) and other statistics for the direct recognition and the re-speaking are introduced in Table 1.

Table 1: Experimental results and statistics.

	Direct recognition	Re-speaking
Recognition WER	25.83 %	4.93 %
Subtitles WER	25.83 %	3.49 %
Words	19 422	8 724
Punctuation marks	0	1 561
OOV	1.99 %	0.03 %
Perplexity	1 706	555

7 CONCLUSIONS

Two methods for live subtitling of TV ice-hockey commentary were presented. The first method is based on the direct recognition of an original audio track with ice-hockey commentary. The second one is so-called "re-speaker" method. In this method a specialized speaker re-speaks the original dialogues.

The problem with the first method was that the recognizer did not work well enough (our previous experiments showed that the WER must be less than 20 % for the subtitles to be understandable). This happens primarily because the speech commentary contains too much noise, whistles, drums, etc. The problem is, unfortunately, difficult to solve because it is not possible to ensure a clean comment since the commentators often sit in the stadium among the audience.

Re-speaking of ice-hockey matches reduces the number of words in final subtitles due to simplifying as described above. Moreover, a perplexity of the task is considerably reduced, because the re-speaker rephrases original incoherent commentary to more comprehensible sentences. The only three out-of-vocabulary (OOV) words in case of re-speaking were added to the recognition system by the re-speaker just during re-speaking. All aforesaid issues result in great word error rate reduction of final subtitles by re-speaking by 86 % relatively.

ACKNOWLEDGEMENTS

This research was supported by the Technology Agency of the Czech Republic, project No. TA01011264.

REFERENCES

Evans, M. J., 2003. Speech Recognition in Assisted and Live Subtitling for Television. *WHP 065*. BBC R&D White Papers.

Trmal, J., Pražák, A., Loose, Z., Psutka, J., 2010. Online TV Captioning of Czech Parliamentary Sessions. *Lecture Notes in Computer Science*, Springer, Heidelberg, Volume 6231.

Pražák, A., Loose, Z., Psutka, J., Radová, V., 2011. Four-phase Re-speaker Training System. In *SIGMAP, International Conference on Signal Processing and Multimedia Applications*.

Skorkovská, L., Ircing, P., Pražák, A., Lehečka, J., 2011. Automatic Topic Identification for Large Scale Language Modeling Data Filtering. *Lecture Notes in Computer Science*, Springer, Heidelberg, Volume 6836.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, Volume 87.

Koehler, J., Morgan, N., Hermansky, H., Hirsch, H. G., Tong, G., 1994. Integrating RASTA-PLP into speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*.

Psutka, J, Psutka, J. V., Ircing, P., Hoidekr, J., 2003. Recognition of spontaneously pronounced TV ice-hockey commentary. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Psutka, J. V., 2007. Robust PLP-Based Parameterization for ASR Systems. In *SPECOM, International Conference on Speech and Computer*.

Zajíc, Z., Machlica, L., Müller, L., 2009. Refinement Approach for Adaptation Based on Combination of MAP and fMLLR. *Lecture Notes in Computer Science*, Springer, Heidelberg, Volume 5729.

Romero-Fresco, P., 2009. More haste less speed: Edited versus verbatim respoken subtitles. *Vigo International Journal of Applied Linguistics*, University of Vigo, Number 6.