# Captioning of Live TV Commentaries from the Olympic Games in Sochi: Some Interesting Insights*

Josef V. Psutka, Aleš Pražák, Josef Psutka, and Vlasta Radová

Department of Cybernetics, West Bohemia University, Pilsen, Czech Republic.
psutka_j,aprazak,psutka,radova@kky.zcu.cz

**Abstract.** In this paper, we describe our effort and some interesting insights obtained during captioning more than 70 hours of live TV broadcasts from the Olympic Games in Sochi. The closed captioning was prepared for ČT Sport, the sport channel of the public service broadcaster in the Czech Republic. We will briefly discuss our solution for distributed captioning architecture on live TV programs using re-speaking approach as well as several modifications of existing live captioning application (especially LVCSR system), but also the way of re-speaking of a real TV commentary for individual sports. We will show that a re-speaker after hard training can achieve such accuracy (more than 98 %) and readability of captions which clearly outperform accuracy of captions created by automatic recognition of TV soundtrack.

**Keywords:** live captioning, speech recognition, re-speaking

## 1 Introduction

One very interesting application area of an automatic speech recognition is the captioning of live TV programs [2,3,6,9]. There are basically two approaches to this task – automatic recognition of a TV soundtrack or human re-speaker who listens to the TV commentary and re-speaks it to the LVCSR system. Both methods have their advantages/disadvantages and are suitable for different types of TV genres and programs. The subtitling from a TV soundtrack is cheaper, but it is more prone to errors that can lead to the creation of obscure subtitles. The choice of a suitable method may also be influenced by the level of background noise, the possibility of overlapping speakers, a manner of their speech, expected frequency of OOV words (even if the vocabulary and language model are tuned to a given genre), etc.

Our research group has extensive experience with both methods of subtitling. During the past years we have subtitled for the Czech TV hundreds of hours of TV broadcasts of meetings from both chambers of the Parliament of the Czech republic (by processing of the TV soundtrack) [8], as well as live programs, such as political debates, entertainment shows and also sports events (using a re-speaker) [12].

The theme of this article was inspired by recent subtitling more than 70 hours of live broadcasts from the Olympic Games in Sochi. We subtitled selected broadcasts of

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

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 515–522, 2014.

ice-hockey matches, speed skating, figure skating, biathlon, alpine skiing, ski jumping, cross-country skiing and also opening and closing ceremonies for ČT4 Sport channel. Everything was done using re-speakers who employed our in-house LVCSR system equipped with specific vocabularies and language models specially prepared for each sport discipline.

At the end of the Olympic competitions, we wanted to answer the question whether it would be possible to provide subtitles for live broadcasts of some sports directly by the processing of accompanying soundtrack, i.e. by recognition of TV commentaries. It is evident that such method of subtitling is cheaper and if it achieved a satisfactory accuracy and intelligibility of generated subtitles, it could be used for some sports during future broadcasts. For this reason we performed a series of experiments where we compared the results of subtitling real broadcasts from Sochi (with our re-speakers) and those that would be obtained if we used directly the automatic recognition of TV commentaries. The results of these experiments, including a brief description of our LVCSR system are discussed in the next sections of this article.

## 2 Acoustic Modeling

### 2.1 Acoustic model for recognition of original soundtrack

For recognition of speech of TV commentators directly from the soundtrack we had to develop a special acoustic model. The acoustic data was collected over several years, especially from TV broadcasts of the Ice-hockey World Championships, the FIFA World Cups and also the last Winter Olympic Games in Vancouver. All these sports events were broadcasted by the Czech television and the total amount of data was more than 100 hours of speech.

The PLP parameterization was used in the front-end module. The sampling frequency was 22 kHz and we used 27 band pass filters and 16 cepstral coefficients with delta and delta-delta features. Due to a very intense background noise, many noise reduction techniques were tested and compared (more details can be found in [4]). The best recognition results were achieved using J-Rasta techniques [1].

The basic speech unit in all our experiments was a three-state HMM with a continuous output probability density function assigned to each state. Phonetic decision trees were utilized to tie states of the triphones. Various experiments were done to ascertain the best recognition results depending on the number of clustered states and also on the number of mixtures per state. The best setting from the recognition point of view was 32 mixtures of multivariate Gaussians for each of 3018 states (see [5] for methodology).

### 2.2 Acoustic model for re-speaking

The main objective of a re-speaker is to take heed to the original speakers and re-speak their dialogues. Unlike the direct recognition of the original soundtrack, a re-speaker can achieve higher transcription accuracy and create easily readable subtitles. An equally important task is simplifying of subtitles, if appropriate, to achieve greater intelligibility.

There are many people among the target audience (for example elderly people), who have limited reading speed capability and subtitles with more than 180 words per minute are frustrating for them [7]. This simplification of subtitles or their re-formulation are in case of direct soundtrack recognition yet unsolvable problem. Similar problems appear 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 capable to keep up.

Each re-speaker has his/her own personal acoustic model. The front-end consisted of PLP features (19 filters, 100 frames per second) with delta and delta-delta coefficients followed by a cepstral mean subtraction. The total dimension of feature vectors was 36. Similar as in a direct recognition system the basic speech unit is a clustered three-state HMM with a continuous output probability density function assigned to each of 4,922 states. The number of mixtures depends on the specific speaker and it is about 22 mixtures of Gaussians per state. Of course, it is assumed that the speech of a re-speaker will be uttered clearly in a quiet environment.

In case of misrecognition, a re-speaker is able to erase the subtitle by a keyboard command and re-speaks it again. Other keyboard commands are used for punctuation marks and original speaker coloring (details can be found in [12]). The job of a re-speaker is quite difficult and requires special long-term training [10]. Unfortunately, not all re-speaker candidates possess the necessary skills. In our experience, a skilled re-speaker can handle maximally one to two hours of subtitling without a pause. A more detailed description of the training of re-speakers can be found in [12].

## 3 Vocabularies and Language Models for Individual Sports

Although the application of automatic speech recognition technology is a better solution for live captioning that surpasses the manual transcription using the keyboard, stenotype or velotype, it still causes some issues that result from the natural limitations of the recognition system. For example, it may be expected that even with very large vocabulary (we use LVCSR with more than one million words), some words needed during live captioning will always be out-of-vocabulary. The re-speaker can use his hands to write the new word or even to add it to the recognition system directly during re-speaking. However, if the vocabulary is too sparse, this may have considerable impact on quality and delay of final captions.

Since each sport has its specific terms and expressions that are commonly used during TV commentary of the sport, the best way is to use the transcriptions of the commentary of the given sport for vocabulary preparation and language model training. In addition, the transcriptions contain the names of sportsmen, their nationalities and teams involved in the match or competition. Even when these names are added to the recognition system they cannot embrace all the names of sportsmen in the matches or competitions that will be captioned in the future. This problem leads to a class-based language model, where its classes should be filled before each live captioning.

We did manual transcriptions of TV commentaries of nine sports participating in the Olympic Games in Sochi. The commentaries of the last Olympic Games in Vancouver were used. To make the corpus independent of specific Olympic Games and their

participants, each name of player, competitor, team, nationality or sport place was labeled. The names which did not relate to the transcribed match or competition were not labeled, because the commentators may use them freely (for example legendary sportsmen – "Bjoerndalen", "Plyushchenko", "Jagr" etc.). Different labels were used for the names of players and competitors, for the names of competing teams or countries and participant nationalities and for the designations of sport places. In addition, each label was supplemented by a number representing one of 6 (excluding vocative) grammatical cases of the expression.

Later on, during language model training, first two labels were automatically divided into basic and possessive forms based on the word ending. Finally, taking into account the above mentioned labels instead of the individual words, six class-based language models, each with 30 classes (12 classes for players and competitors, 12 classes for teams and nationalities and 6 classes for sport places) were trained. These language models comprise also some punctuation marks (comma, colon, full stop and question mark) indicated by the re-speaker during re-speaking. Counts of tokens and labels of each type in the training corpus for individual sports are reported in Table 1.

**Table 1.** Statistics of the training corpus

| | **# of tokens** | **# of players & competitors** | **# of teams & nationalities** | **# of sport places** |
|---|---|---|---|---|
| Alpine skiing | 119,191 | 4,413 | 1,293 | 554 |
| Biathlon | 68,960 | 4,085 | 1,400 | 269 |
| Cross-country skiing, ski jumping, Nordic combined | 227,680 | 11,113 | 4,362 | 1,141 |
| Figure skating | 86,395 | 2,170 | 876 | 201 |
| Ice-hockey | 922,694 | 89,870 | 24,790 | 1,047 |
| Speed skating, short track | 126,233 | 4,953 | 2,590 | 530 |

Even complete transcriptions of TV commentaries from the last Olympic Games in Vancouver are not sufficient for robust language modeling with large vocabularies. That is why we used other language models based on the data from the sport domain (9M tokens from TV news transcriptions, 88M tokens from newspapers and 94M tokens from internet news) and mixed them with the class-based language models for individual sports. The weights of individual language models were set to minimize the perplexity on the transcriptions referred in Table 1 by the cross-validation technique.

While each Olympic sport has its own fans among deaf and hard-of-hearing, the opening and closing ceremonies of the Olympic Games are watched by the most viewers. At the same time, it is the most challenging task for live captioning, because the scope of the commentary is very wide and the problem of sparse vocabulary is even more significant. Unfortunately, we did not receive any detailed information about the ceremonies ahead of their captioning. While we covered the names of legendary sportsmen and sportsmen participating in the Olympic Games by using class-based

language models created for individual sports, names of artists, songs and work of arts were covered only partially.

To prepare the best language model for captioning of opening and closing ceremonies of the Olympic Games, we used other data from non-sport domains, some texts concerning Russian history and the transcriptions of ceremonies from several Olympic Games as well (see [11] for more detailed information on the selection of appropriate data). Anyway, the largest part of live captioning was made by the re-speaker who has to overcome the imperfectness of the recognition system with the vocabulary containing over 1.1M words. For example, in the case of repeated out-of-vocabulary word he can simply (on-line) add such new word to the recognition system by typing it.

## 4 Experimental Results

The Olympic Games were broadcasted on two channels of the Czech TV. Although we captioned only some live transmissions chosen by the Czech Television, some days we had to employ all six skilled re-speakers who alternated every two hours. Totally we captioned 74 hours of live transmissions during two weeks. The live captioning was performed by specially trained re-speakers from their homes connected through the internet and ISDN network to the Czech Television [12].

As the described class-based language models specifically trained for each sport were used, the re-speaker had to fill the classes before each live captioning. We enhanced our captioning software for a class management. Before each live captioning session the re-speaker adds the names and surnames of sportsmen participating in the match or competition based on official start lists on the web pages. Since the pronunciation of foreign name depends on its original language, the re-speaker enters a pseudotranscription (a transcription using Czech letters) for each name, which is processed by common rule-based grapheme-to-phoneme conversion. We do not know the exact pronunciation that will be used by a TV commentator, so we make the best guess based on our knowledge of foreign languages and if the pronunciation then varies, the re-speaker translates the name to the right form demanded by the recognition system. Twelve language model classes are filled with names and surnames automatically declined (based on indicated sex) to 6 basic grammatical cases and 6 grammatical cases of their possessive forms. Both a surname only and a combination of name and surname are generated. Since the declension of foreign names is a very complex problem, the re-speaker may check and correct all generated items. These items are then added to the language model classes based on their grammatical case and form. The same process is used for the names of teams, nationalities and sport places. All items may be prepared in advanced, so the re-speaker only chooses predefined selection, items are added to the language model of the recognition system (including all class n-grams) and the re-speaker immediately starts captioning.

The live captioning through re-speaking of Winter Olympic sports has some specifics compared to the live captioning of political debates or TV shows. Due to the delay of live captions (about 5 seconds on average) some information may be irrelevant when the captions are displayed to the viewer, for example designations of jumps during figure skating or passes between ice-hockey players. In addition, superfluous captions

may distract viewer's attention from the sport itself. Thus an additional aim of the re-speaker is to filter the information contained in sport commentaries and to deliver to the viewer all the important and interesting information in the form that do not bother nor distract the viewer. In Table 2, you can see the average number of words per minute in original and re-spoken commentaries. The re-speaking factor shows the proportion of re-spoken words against the original comment.

**Table 2.** Re-speaking factor

| | **Source of speech** | | **Re-speaking factor [%]** |
|---|---|---|---|
| | **Original** | **Re-spoken** | |
| | **words per minute** | | |
| Alpine skiing | 107 | 89 | 82.9 |
| Biathlon | 114 | 88 | 77.6 |
| Ceremonies | 81 | 56 | 69.2 |
| Cross-country skiing, ski jumping, Nordic combined | 116 | 91 | 77.8 |
| Figure skating | 50 | 42 | 84.5 |
| Ice-hockey | 91 | 57 | 62.7 |
| Speed skating, short track | 113 | 86 | 76.0 |

As can be seen in Table 2, there are not big differences between re-speaking factors of individual Olympic sports with the exception of ice-hockey and the opening and closing ceremony. The problem with subtitling ice-hockey by re-speaking is (already mentioned above) 5 second delay between the actual commentary and the subtitle displayed to the viewer. Some in-game situations may be irrelevant with such delay. 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, re-speakers concentrate mainly on rich commentary of interesting situations not covered by the video content.

During captioning of the opening and closing ceremonies very difficult tasks had to be solved by re-speakers, because they added to the recognition system many OOV words. These often refilling OOV words, then induced a significant simplification of the re-spoken commentaries due to a lack of time.

One hour of each Olympic sport was manually transcribed both the original as well as the re-speaker soundtrack to evaluate the type of suitable captioning method. The Table 3 shows OOV, perplexity and word error rate (WER) for recognition from the original TV soundtrack and also the WER for recognition from the re-spoken soundtrack. The OOV is not shown in the second case, because it is almost zero due to adding of all OOV words to the recognition system by hand. All recognition experiments were performed with the corresponding class-based language model.

A re-speaker evaluation was performed according to the evaluation scheme proposed especially for captioning through re-speaking introduced in [12]. The first level evaluation is represented by a standard word error rate, while the semantic accuracy represents

**Table 3.** Experimental results

| | Source of speech | | | | |
|---|---|---|---|---|---|
| | Original soundtrack | | | Re-spoken | |
| | OOV [%] | Perplexity | WER [%] | WER [%] | Semantic Acc [%] |
| Alpine skiing | 0.99 | 579 | 32.02 | 1.34 | 93.06 |
| Biathlon | 0.66 | 514 | 34.53 | 1.23 | 91.20 |
| Ceremonies | 3.58 | 910 | 45.77 | 2.83 | 83.50 |
| Cross-country skiing, ski jumping, Nordic combined | 0.90 | 602 | 32.93 | 1.61 | 92.70 |
| Figure skating | 0.87 | 395 | 35.45 | 1.30 | 97.84 |
| Ice-hockey | 0.93 | 545 | 25.97 | 1.03 | 92.77 |
| Speed skating, short track | 0.28 | 315 | 27.66 | 1.14 | 86.29 |

the third level of the evaluation – the 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).

## 5 Conclusion

In this article we described our efforts to build a system for captioning broadcasts of the Olympics Games in Sochi. We discussed not only the LVCSR system, which is equipped with special class-based language models, but also two principal methods for captioning live TV programs, especially sports events. The specially prepared class-based language models reduce significantly the perplexity and OOV and contribute to the robust recognition of key information such as names of players, nationalities, etc. It is evident, according to the obtained results, that we cannot use the recognition of the original soundtrack due to a very high WER. This is caused partly by a very noisy background (cheering, music, whistles, drums, etc.) in the original commentary and partly by a relatively higher OOV (opening ceremony) even if we used 1.1M words in vocabulary.

Such difficulties disappear when we use re-speaking method for captioning. In this method, an experienced and highly trained speaker re-speaks (in a quiet environment) and optionally simplifies the original commentary. Moreover, a perplexity of the task is considerably reduced, because the re-speaker rephrases original often incoherent commentary to more comprehensible sentences. The WER decreased rapidly from 33.5% (direct recognition) to 1.5% (re-speaking) on average.

## References

1. Koehler, J., Morgan, N., Hermansky, H., Hirsch, H.G., Tong, G.: Integrating RASTA-PLP into speech recognition. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing (1994), pp. 421–424 vol. 1

2. Evans, M.J.: Speech Recognition in Assisted and Live Subtitling for Television. R&D White Paper WHP 065, BBC Research & Development (2003)
3. Marks, M.: A distributed live subtitling system. R&D White Paper WHP 070, BBC Research & Development (2003)
4. Psutka, J, Psutka, J.V., Ircing, P., Hoidekr, J.: Recognition of spontaneously pronounced TV ice-hockey commentary. In: ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (2003), pp. 169–172
5. Psutka, J.V.: Robust PLP-Based Parameterization for ASR Systems. In: SPECOM, International Conference on Speech and Computer (2007), pp. 509–515
6. Ortega, A., Garcia, J.E., Miguel, A., Lleida, E.: Real-Time Live Broadcast News Subtitling System for Spanish. In: 10th Annual Conference of the International Speech Communication Association, pp. 2095–2098. Causal Productions (2009), pp. 2095–2098
7. Romero-Fresco, P.: More haste less speed: Edited versus verbatim respoken subtitles. Vigo International Journal of Applied Linguistics, University of Vigo, Number 6. (2009). pp. 109–133
8. Trmal, J., Pražák, A., Loose, Z., Psutka, J.: Online TV Captioning of Czech Parliamentary Sessions. Lecture Notes in Computer Science, Springer, Heidelberg, Volume 6231. (2010). pp 416–422.
9. Bordel, G., Nieto, S., Penagarikano, M., Rodriguez-Fuentes, L.J., Varona, A.: Automatic Subtitling of the Basque Parliament Plenary Sessions Videos. In: 12th Annual Conference of the International Speech Communication Association, pp. 1613–1616. Causal Productions (2011). pp. 1613–1616
10. Pražák, A., Loose, Z., Psutka, J., Radová, V.: Four-phase Re-speaker Training System. In: SIGMAP, International Conference on Signal Processing and Multimedia Applications (2011). pp. 217–220
11. Švec, J., Hoidekr, J., Soutner, D., Vavruška, J.: Web text data mining for building large scale language modelling corpus. In: Habernal, I., Matoušek, V. (eds.) Text, Speech and Dialogue, LNCS, vol. 6836, pp. 356–363. Springer, Heidelberg (2011)
12. Pražák, A., Loose, Z., Trmal, J., Psutka, J.V., Psutka, J.: Novel Approach to Live Captioning Through Re-speaking: Tailoring Speech Recognition to Re-speaker's Needs. In: 13th Annual Conference of the International Speech Communication Association, pp. 1370–1373. Red Hook: Curran Associates, Inc. (2012)