Online TV captioning of Czech Parliamentary Sessions *

Jan Trmal[1], Aleš Pražák[2], Zdeněk Loose[1], and Josef Psutka[1]
{jtrmal, zloose, psutka}@kky.zcu.cz
{ales.prazak}@speechtech.cz

[1] Department of Cybernetics, University of West Bohemia,
Plzen, Czech Republic
[2] SpeechTech, s.r.o, Plzen, Czech Republic

Abstract. In the paper we introduce the on-line captioning system developed by our teams and used by the Czech Television (CTV), the public service broadcaster in the Czech Republic.
The research project is targeted at incorporation of speech technologies into the CTV environment. One of the key missions is the development of captioning system supporting captioning of a "live" acoustic track. It can be either the real audio stream or the audio stream produced by a shadow speaker. Another key mission is to develop software tools and techniques usable for training the shadow speakers.
During the initial phases of the project we concluded that the broadcasting of the Parliamentary meetings of the Chamber of Deputies fulfills the necessary conditions that enable it to be captioned without the aid of the shadow speaker. We developed a fully automatic captioning pilot system making the broadcasting of Parliamentary meetings of the Chamber of Deputies accessible to the hearing impaired viewers.
The pilot run enabled us and our partners in the Czech TV to develop and evaluate the complete captioning infrastructure and collect, review and possibly implement opinions and suggestions of the targeted audience.
This paper presents our experience gathered during first years of the project to the public audience.

1 Introduction

The first application of ASR (Automatic Speech Recognition) system for real-time captioning of live broadcasts has arguably been announced by BBC in 2003[1]. Since then, similar systems have been developed and employed in production use in several countries all around the world (see for example [2], [3], [4]). However, the speech recognition done on real world acoustic track is much more difficult than the recognition accuracies presented in vast majority of the scientific papers would suggest – consider speaker dialects and speaker emotional

* This research was supported by the Ministry of Education of the Czech Republic, project No. 2C06020 and by the University of West Bohemia, project No. SGS-2010-054.

states, necessity of very large vocabularies for highly inflectional languages, diverse acoustic environments and large variety of signal distortions introduced by different technical equipments used for acoustic signal transport, storage and processing. To overcome majority of such problems, an alternative approach, called "shadow speaker" (or re-speak) approach is often employed. The principle of such approach is as follows. Instead of generating of the captions from the real-world acoustic track, an indirect approach is used. The real-world track is listened to and re-spoken by a skilled and specifically trained employee – the shadow speaker.

This simplifies the task of ASR significantly – the shadow speaker works in a quiet environment, uses a well defined acoustic channel and is not under an emotional stress. Moreover, the acoustic model as well as language model in the used ASR system can be tuned specifically for the given speaker. On the top of it, one or more human correctors usually correct misrecognized words, add punctuation, perform hyphenation (if needed) and format the recognized text to the final captions. With this setup, the reported accuracies are usually highly over 95 %.

As already said, one of the main objectives was development of a captioning system supporting real-time online operation, generating captions either from the real acoustic track or from the track respoken by the shadow speaker. Since there were no skilled shadow speakers available in the beginning of the project, we have identified several TV shows that we found captionable using the real acoustic track. The set of suitable shows contained weather news, specific discussion shows and meetings of the Chamber of Deputies and the Senate of the Parliament of the Czech Republic. After discussion with representatives of the CTV, we decided to pursue captioning of the meetings of the Chamber of Deputies of the Czech Parliament.

The Chamber of Deputies has 200 members, elected every four years. Although the number of members is quite large, only a small portion of them acts actively during the meetings. Moreover, a large portion of the active speakers stays in the service several electoral terms. These speakers are skilled orators and the rules of procedure enforce that no member may speak unless called upon to do so by the chairman. The audience chamber has a professional, high quality audio capture system and the acoustic channel is stable and the quality is sufficient for ASR.

2 Captioning system architecture

Because of the strict security policies at the premises of Czech Television, the system was designed as highly distributed, see Fig. 1. The interconnection between CTV and UWB is done by a point-to-point connection over the ISDN network. However, instead of using the ISDN voice services, the ISDN is used only as data carrier. The two B channels of the ISDN-BRI are bonded, providing bandwidth of 128 kbit/s.

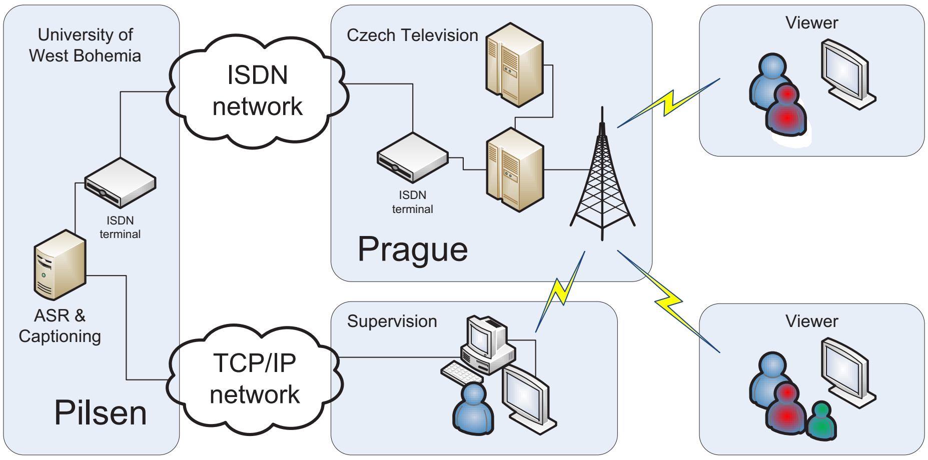


Fig. 1. Geographic scheme of the pilot captioning system

To provide transparent interconnection over the ISDN network, specialized terminal adapters CDQPrima 210 are used. Outside of the security consideration, this design decision helped us to isolate ourselves from the network issues and audio streaming issues. The resulting formatted captions are sent over a serial line back to the codec and then over the ISDN network using the "ancillary data" feature of the codecs. Then they are broadcasted as EIA-608 (line 21) captions available using teletext page 888.

The ISDN audio codecs support quite a large variety of audio compression standards. We evaluated several supported compression standards: MPEG-1 Audio Layer II, MPEG-1 Audio Layer III and G.722. We evaluated the suitability of each of this codecs by transcoding the training audio data by a software implementation of the specific codec, training the acoustic models and performing a recognition tests on the heldout data. Because of the recognizer accuracy, we have originally chosen the MPEG-1 Audio Layer III standard however because of technical reasons not tied to, the 128 kbit/s, 48 kHz MPEG-1 Audio Layer II (MP2) codec is used in the current pilot system. The recognizer accuracy drop is not fatal and the MP2 standard offers lower algorithmic delay.

The time delay from word utterance to receiving the data by broadcast receivers is about 2–4 seconds. For the detailed analysis, see Fig. 2. It is clear, that the most prominent factor of the delay is defined by the caption formatting and postprocessing subsystem. However, this delay is dictated by the length of subtitle itself, since the caption is formed from the recognized text spanning the mentioned time interval.

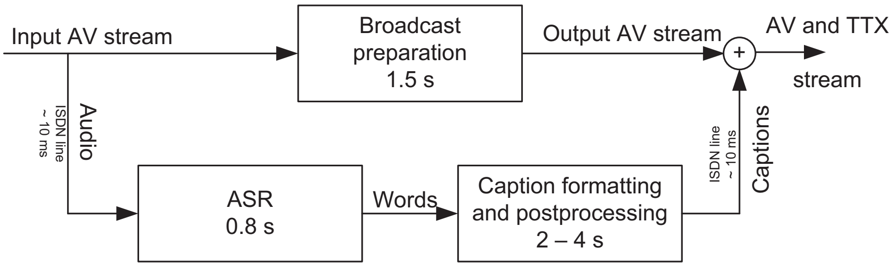


Fig. 2. Timing scheme of the pilot captioning system

3 Speech Recognition

During the preparatory phase of the project, we gathered about 200 hours of parliamentary sessions recordings. The training data were obtained by means of recording the broadcasting. We use three-state HMMs and 8 mixtures of multivariate Gaussians for each state. The total number of approx. 50k Gaussians is used for the speaker-independent model. In addition, discriminative training techniques were used (see [5]). The acoustic model were trained on 44.1 kHz audio stream using PLP parametrization with 19 PLP filters and 12 PLP cepstral coefficients, augmented by delta and delta-delta parameters. To model the transport channel the source data were transcoded (encoded and decoded again) using a software implementation of the 128 kbps MPEG-1 Audio Layer II compression codec.

For language modeling we used the stenographic transcripts that are made public by law. To allow subtitling of arbitrary (including future) electoral period, five classes for representative names in all grammatical cases were created. See [6] and [7] for details. The vocabulary size is approx. 200k words. For the fast online recognition, we use a class-based bigram language model with modified Knesser-Ney discounting trained by SRI Language Modeling Toolkit. For a more accurate confidence measure of recognized words, the class-based trigram language model is used.

Before each captioning session, the language model is adapted using public materials from world-wide web – we use the related texts to integrate the new words and the related n-grams.

The speech recognition accuracy is variable, depending on the type of procedure, orating skills, states of mind and tempers of the speakers and of course the topic. In overall, it is about 85 %–88 %. When the flow of the meeting is highly directed by the rules of procedure (for example voting), the recognition accuracy is over 90 %. The majority of the recognition errors is caused by missing, redundant or misplaced prepositions – according to our experience this kind of error does not hamper the legibility of the subtitles. Moreover, some of these errors can be corrected by postprocessing.

3.1 Automatic caption generation

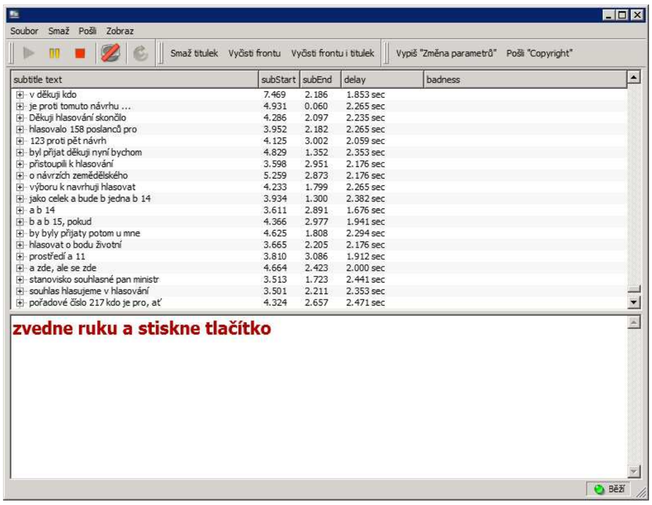
We developed two versions of caption formatting and postprocessing (CFP) subsystem. The first version was developed as a proof-of-concept. It was single user application only, the user had to use the RDP protocol client to log on the captioning server and the possibilities of user assisted caption editation or corrections were limited. The listening-in feature was also accomplished by the RDP client-server architecture.

The development of the second version of the CFP subsystem started after the pilot system evaluation (see below). We have designed a fully distributed client-server application that takes the comments and suggestions of the interviewees into account as well.

The server is the computer where the ASR and the CFP run. It supports automated operation without intervention of any kind. The client is any computer that runs the client software. The client software supports the control of the server (starting, stopping, pausing of the recognition) and editation of the formatted captions.

The number of client users is unlimited. The communication protocol runs over TCP/IP and is secured via SSL. The clients authenticate themselves using an encrypted client SSL certificates issued by the internal certification authority. This is to ensure secure operation independent on the location of the client.

The communication protocol is proprietary, message oriented. To support the listening-in feature even over low-bandwidth internet connection, the server compress the original acoustic track using the Speex codec and streams the compressed track instead of the original track.



subtitle text	subStart	subEnd	delay	badness
v děkuji kdo	7.469	2.186	1.853 sec	
je proti tomuto návrhu ...	4.931	0.060	2.265 sec	
Děkuji hlasování skončilo	4.286	2.097	2.235 sec	
hlasovalo 158 poslanců pro	3.952	2.182	2.265 sec	
123 proti pět návrh	4.125	3.002	2.059 sec	
byl přijat děkuji nyní bychom	4.829	1.352	2.353 sec	
přistoupili k hlasování	3.598	2.951	2.176 sec	
o návrzích zemědělského	5.259	2.873	2.176 sec	
výboru k navrhuji hlasovat	4.233	1.799	2.265 sec	
jako celek a bude b jedna b 14	3.934	1.300	2.382 sec	
a b 14	3.611	2.891	1.676 sec	
b a b 15, pokud	4.366	2.977	1.941 sec	
by byly přijaty potom u mne	4.625	1.808	2.294 sec	
hlasovat o bodu životní	3.665	2.205	2.176 sec	
prostředí a 11	3.810	3.086	1.912 sec	
a zde, ale se zde	4.664	2.423	2.000 sec	
stanovisko souhlasné pan ministr	3.513	1.723	2.441 sec	
souhlas hlasujeme v hlasování	3.501	2.211	2.353 sec	
pořadové číslo 217 kdo je pro, ať	4.324	2.657	2.471 sec	

zvedne ruku a stiskne tlačítko

Fig. 3. Main screen of the subtitling software

Caption types The standard EIA-608([8]) defines three kinds of closed captions:

- Roll-up – the words appear one by one until they fill the line, then the line is rolled up (possibly erasing the line on very top) and the new line is started to be filled. This is the recommended format of captions for live events, where the captions are generated on-the-fly.
- Pop-on, block – the captions are pre-formatted to span one or several lines and appear instantaneously, erasing the previous captions block. This kind of captions is used prevailingly nowadays.
- Paint-on – the captions appear on the screen by letter-by-letter fashion, forming a stationary block in the end, just like the pop-on captions. Almost unused nowadays.

Because of CTV policies, the captions formatting subsystem we developed produces the pop-on captions. By custom, the captions are displayed as block spanning one or two lines, depending on speaker speech cadence.

To produce eye-pleasing captions we developed a special formatting algorithm based on dynamic programming. The algorithm ranks all the possible variants of captions generated from the given block of recognizer results according to their amenity and chooses the best. If even the best caption scores lover than a certain threshold, the caption is not produced that time. There are several factors that are taken into account when evaluating the amenity:

- delay from the previous caption – too short or too long delays are penalized
- ratio of line filling – shorter lines are penalized
- ratio of line lengths – evaluated only in case where the proposed caption spans two lines, to obtain lines of approximately the same size
- line breaking rules – caption line is not allowed to end with preposition
- internal CTV policies
- ad-hoc rules

Moreover, the caption variant formatting process performs rule-based editation of the resulting caption. During this phase, punctuation is filled in and some kinds of recognizer errors and speaker tics are corrected. The corrected recognizer errors include stray prepositions, multiple prepositions, the corrected speaker tics consisting of word repetitions, hesitation sounds, etc.

In the pilot version of the captioning system, the editation rules were produced by an expert, however, we are currently working on an automatic and a semi-automatic inference of the editation rules.

4 Evaluation of the system among the targeted audience

Several months after the pilot captioning system was deployed, an evaluation of the system was arranged. The questionnaires consisted of 7 questions formed into a "single choice from multiple choices" form available online on the web.

1. Do you think the project shall continue?

2. If you have seen the captioned broadcast, do you think the captions were understandable?
3. Do you see the errors in punctuation as a big problem?
4. Did the delay hamper the viewing experience?
5. Do you think that even a partial correction of the biggest errors is beneficial?
6. What kind of captions would you prefer?
7. What kind of speaker change notification would you prefer?

Moreover, the interviewees were asked to append their own comments or questions. We find that the evaluation was necessary to obtain the perspective of the potential end-users.

The findings were as following. The majority of the responders found the subtitles comprehensible and does not find the 2–5 seconds delay to be a problem. The same holds for the punctuation and for the error correction − responders do not feel any of it is vital. Note that this is most probably because of the nature of the programme. Very important were the answers to the last two questions. A majority of the responders prefer the speaker change notification to be accomplished via caption color change and is in favor of the roll-up subtitles. From the additional comments it is clear that the most problematic aspect of the subtitling is combination of the rolling realtime information line that is overlayed over the original program during the broadcast preparation in CTV. The responders found it very uncomfortable, since even without the infoline they have to split their attention between the original video and the subtitles and scrolling of the infoline disturbs the viewers' attention. This can be alleviated by rendering a full-size black box around the subtitles. Lastly, one comment suggested the automatically captioned subtitles should use different pictogram than the programmes with manual captions.

5 Conclusion and Future Prospects

In the paper we presented an overview of a captioning system used by a public service broadcaster in the Czech Republic. The captioning system supports real-time online captioning either from the real acoustic track or the acoustic track produced by the shadow speaker. The system is fully functional; however we plan to enhance it in several ways. Besides improvements of the language model adaptation process and automatic switching between gender-specific and speaker-dependent acoustic models we aim to simplify and possibly eliminate the necessity of human assisted caption correction.

This includes generation of automatic correction rules from large text base, possibly using methods of automatic translation. The semantic punctuation can be supplemented by means enhancing the automatic correction subsystem to use prosodic and acoustic features.

References

1. Evans, M.J.: Speech Recognition in Assisted and Live Subtitling for Television. White Paper WHP065, BBC Research & Development (September 2003)

2. Homma, S., Kobayashi, A., Oku, T., Sato, S., Imai, T., Takagi, T.: New real-time closed-captioning system for japanese broadcast news programs. (2008) 651–654
3. Saraclar, M., Riley, M., Bocchieri, E., Goffin, V.: Towards automatic closed captioning: Low latency real time broadcast news transcription. (2002)
4. Boulianne, G., Beaumont, J., Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., Comeau, M., Ouellet, P., Osterrath, F.: Computer-assisted closed-captioning of live TV broadcasts in French. In: Ninth International Conference on Spoken Language Processing. (2006)
5. Povey, D.: Discriminative Training for Large Vocabulary Speech Recognition. PhD thesis, Cambridge University, Engineering Department (2003)
6. Pražák, A., Müller, L., Psutka, J.V., Psutka, J.: LIVE TV SUBTITLING - fast 2-pass LVCSR system for online subtitling. In: SIGMAP 2007, International Conference on Signal Processing and Multimedia Applications. (2007)
7. Pražák, A., Psutka, J., Hoidekr, J., Kanis, J., Müller, L., Psutka, J.: Automatic online subtitling of the czech parliament meetings. Lecture Notes in Artificial Intelligence (2006) 501–508
8. EIA-608-B: Recommended practice for line 21 data service. Technical Report EIA/ANSI 608-B, Electronic Industries Alliance (September 1994)