Implementation of a Live Dialectal Media Subtitling System

Michael Stadtschnitzer, Christoph Schmidt

Fraunhofer IAIS, Schloss Birlinghoven, Sankt Augustin, Germany

{Michael.Stadtschnitzer,Christoph.Andreas.Schmidt}@iais.fraunhofer.de

Abstract

Subtitling is a useful technique to fulfil the information needs of deaf and hearing impaired people. Live subtitling is needed especially for live events and is not restricted to television, but can also be provided to persons on site, e.g. to a deaf politician during a parliamentary debate. Live subtitling is demanding since the audio information has to be transformed into text within a few seconds. In this demonstration we present the Fraunhofer IAIS audio and video live subtitling system for standard German and Bavarian. The system was developed in the "Live-Caption" project and consists of an online speech recognition and speaker diarisation system. We employ our entire large annotated standard German broadcast corpus for training. In addition, we apply the system to Bavarian dialect by adapting the acoustic models and the pronunciation lexicon, exploiting a Bavarian media corpus. Due to the real-time restrictions and the spontaneous character of dialectal speech, the system performance is far from perfect, but encouraging.

Index Terms: speech recognition, live subtitling, dialect

1. Introduction

The subtitling of media is a vital means of assisting deaf or hearing impaired people to fulfill their information needs. In recent years, national legislation has forced broadcasters to provide subtitles for a certain quota of their broadcast media. Moreover, subtitles are useful at loud crowded places where the media can be seen but not necessarily be heard, e.g. in restaurants or train stations. Currently, subtitles are usually produced manually or semi-automatically. Besides the manual typing of the spoken text by a human, a common subtitling method is re-speaking, in which a trained person clearly re-speaks the words he/she hears into a microphone, also speaking punctuation marks (e.g. "full stop"). This clean audio can then be fed into a speaker-dependent speech recognition system which produces a recognition result with a very high quality. In contrast to this, the system presented in this work has to deal with different speakers as well as noisy and changing acoustic conditions.

Live subtitling is necessary for the broadcasting of live events such as sport matches, political events or the like. However, subtitles can not only be provided to televisions in the home, but also to screens on site to assist deaf and hard of hearing people. For example, when a deaf politician takes part in a parliamentary session, they might want to read the words spoken by another politician as a live transcript, which allows them to jump back and read a previous sentence. Such live transcription services are currently provided manually by the project partner VerbaVoice in several German federate state parliaments. In this scenario, the task to develop an automatic live subtitling system is aggravated by dialectal speech, since especially in emotional debates some politicians switch from standard German to their local dialect. In addition to our standard German system, we therefore adapt our models to the dialectal situation. For this demo, we adapt our models to the Bavarian dialect into our live subtitling system, leaving other dialects for future work.

Figure 1: Bavarian subtitling, "Dahoam is Dahoam", Episode 1443, Source: Bayerischer Rundfunk (BR).

2. Related Work

The authors in [1] present an automatic subtitling system for the subtitling of arbitrary videos found on the web. The models are trained on Voxforge data. The system in [2] is developed to assist Spanish TV Broadcasters to comply with policies to subtitle 90 % of their content by 2013. The system covers voice activity detection, speech recognition and alignment, discourse segment detection and speaker diarisation to assist post-editing. These systems work offline, but sometimes live subtitling systems are needed. In [3] a live subtitling system tailored for three different live television environments, namely news programmes, sports and magazine is presented. The system measures the subtitle delays to synchronise media and subtitles before delivering the TV program in a dedicated IPTV channel. A real-time automatic video subtitling system for Spanish News is proposed in [4]. The system consists of a text retrieval module which communicates with a news redaction computer system and a speech-text temporal alignment module based on automatic speech recognition (ASR).

3. System description

3.1. Standard German

In [5] we presented the Fraunhofer IAIS standard German large vocabulary continuous speech recognition system (LVCSR) which is based on the Kaldi toolkit [6]. It was trained on 636 hours of speech data taken from the GerTV1000h German Broadcast Speech Corpus presented in the same paper. In this work we use the whole 1005 hours of training data of the GerTV1000h corpus. The acoustic models of the system are based on deep neural networks (DNN) [7]. We also perform a

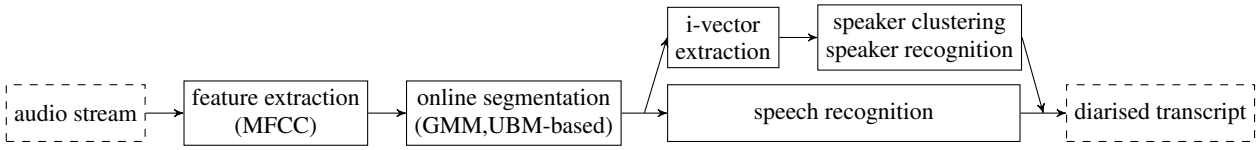

Figure 2: Architecture of the online subtitling system. Speech recognition and speaker segmentation/clustering are done in parallel.

subsequent recurrent neural network (RNN) rescoring approach [8], which is however not part of the online system. The evaluation results of the models on a development set (cf. [5]) and test sets taken from the DiSCo Corpus [9] are shown in Table 1. The results show that the extension of the training material and subsequent RNN rescoring is beneficial.

Model	**Size (h)** Training	**Dev.**	**DiSCo** planned	**DiSCo** spontaneous
DNN [5]	636	22.7	17.4	21.5
DNN	1005	21.3	15.5	19.7
DNN/RNN	1005	20.0	15.3	18.4

Table 1: Results in terms of WER [%] of various configurations

3.2. Bavarian

We noticed that the standard German LVCSR system performance degrades when dialectal speakers, e.g. from Bavaria, Berlin or Cologne are present (e.g. to 90.1 % WER on a Bavarian test set). Hence, we currently tailor the standard German models to Bavarian dialect by adapting the acoustic models to a corpus consisting of about 50 hours of Bavarian broadcast media and by extracting phonetic information which is then used to adapt the pronunciation lexicon. Additionally, the transcripts of the Bavarian corpus are included in the language model.

3.3. Online system

The online speech recognition and speaker diarisation system is based on the Kaldi neural-net-based online decoder. A general system architecture is depicted in Figure 2.

The system input is an audio or video stream. First, an online MFCC feature extraction is applied. This feature stream is pre-segmented by a GMM/UBM-based segmenter, and the segments are provided in parallel to the i-vector extraction / speaker clustering system and the speech recognition system. By comparing the extracted i-vector of the current segment to previous segments, segments of the same speaker can be clustered. In the subtitling system, identical speakers are marked with a certain subtitle colour. In the end, the segment and speaker information is merged with the spoken words obtained from the speech recognition system to create a diarised transcript which can be displayed as a colour-coded subtitle. The online subtitling system provides live subtitles as depicted in Figure 1.

4. Discussion / Conclusions

In this work, we presented a live subtitling system consisting of a speaker diarisation and a speech recognition system with models both for standard German and Bavarian dialect.

The main difference between the online speech recognition system presented in this paper and a standard offline recognition system is the segmentation and speaker clustering system. While in the batch processing mode such a system can find globally optimal segment boundaries and speaker labels, an online system can only rely on the current and past information. Consequently, the system provides a suboptimal speaker clustering, especially when the acoustic conditions change between different shots. To run the system under real-time conditions, we only had to minimally prune the language model obtained from the training data. Thus the degradation compared to an offline system is negligible.

5. Acknowledgements

This work has been partly funded by the German Federal Ministry of Education and Research (BMBF) “KMU-innovativ” funding initiative (funding no. 01IS14023C) for the research project “LiveCaption”.

6. References

[1] A. Mathur, T. Saxena, and R. Krishnamurthi, “Generating Subtitles Automatically using Audio Extraction and Speech Recognition,” in *Proc. of IEEE International Conference on Computational Intelligence and Communication Technology*, Ghaziabad, India, 2015, pp. 621–626.

[2] A. Alvarez, A. del Pozo, and A. Arruti, “APyCA: Towards the Automatic Subtitling of Television Content in Spanish,” in *Proc. of International Multiconference on Computer Science and Information Technology(IMCSIT)*, Wisla, 2010, pp. 567–574.

[3] M. de Castro, D. Carrero, L. Puente, and B. Ruiz, “Real-time Subtitle Synchronization in Live Television Programs,” in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Nuremberg, 2011, pp. 1–6.

[4] J. E. Garcia, A. Ortega, E. Lleida, T. Lozano, E. Bernues, and D. Sanchez, “Audio and Text Synchronization for TV news Subtitling based on Automatic Speech Recognition,” in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Bilbao, 2009, pp. 1–6.

[5] M. Stadtschnitzer, J. Schwenninger, D. Stein, and J. Koehler, “Exploiting the Large-Scale German Broadcast Corpus to boost the Fraunhofer IAIS Speech Recognition System,” in *Proc. Language Resources and Evaluation Conference (LREC)*, Reykjavik, Island, May 2014.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi Speech Recognition Toolkit,” in *Proc. Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2011.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. Audio Speech Lang. Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[8] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Cernocky, “RNNLM - Recurrent Neural Network Language Modeling Toolkit,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[9] D. Baum, D. Schneider, R. Bardeli, J. Schwenninger, B. Samlowski, T. Winkler, and J. Köhler, “DiSCo — A German Evaluation Corpus for Challenging Problems in the Broadcast Domain,” in *Proc. Seventh conference on International Language Resources and Evaluation (LREC)*, Valletta, Malta, may 2010.