Improving Automatic Speech Recognition for Effective Topic Segmentation

Michael Stadtschnitzer, Joachim Koehler, Daniel Stein

Fraunhofer IAIS, 53757 Sankt Augustin, Germany, Email: name.surname@iais.fraunhofer.de

Abstract

Since the last few years a paradigm change takes place in the training of acoustical models for automatic speech recognition (ASR). Hidden Markov models (HMMs) with state-dependent Gaussian Mixture Models (GMMs), which have been the de facto standard for several decades, are now replaced by modern training algorithms i.e. Deep Neural Networks (DNNs) and Subspace Gaussian Mixture Models (SGMMs).

In this work we replace the acoustical model of the Fraunhofer IAIS German broadcast ASR system based on GMM-HMMs by both DNN and SGMM based models and provide and evaluation of the large improvements compared to the results of the baseline ASR system, which is already described and evaluated in previous publications. We further increase the amount of data for training of the acoustical models from 105 to 636 hours and evaluate the improvements made by largely increasing training data size. We also describe our 1,000 h German broadcast speech corpus from which the training data was taken and which will be fully exploited for acoustical model training in the near future.

When making large heterogenous audio-visual broadcast databases searchable, robust ASR provides amongst others a basic technology for subsequent tasks e.g. topic segmentation and topic labelling. We briefly introduce our topic segmentation and labelling system and motivate the importance of employing high quality ASR.

Introduction and Motivation

The improvement of large vocabulary continuous speech recognition (LVCSR) systems is still an ongoing research field in speech technology. The word error rate (WER) for spontaneous speech of broadcast recordings for modern ASR systems varies between approximately 15 to 50 percent at a rough guess, depending on various factors like the language, speaking style, dialects and accents and the quality of input speech data. Especially when the ASR output is further processed for tasks like topic segmentation, topic labelling or named entity recognition it is important to use a modern robust ASR system with WERs below 40 percent absolute in difficult situations to make sense at all. The LVCSR system which we published in [1] does not fulfill this requirement. In this paper we describe our strategies which improved our broadcast German LVCSR system. First, we designed a novel German broadcast speech corpus consisting of 1,000 hours of transcribed speech data recorded from a variety of German broadcast formats which we employ for ASR model training. Second, we employ recent advanced methods for acoustic modelling, i.e. DNNs [3, 2] and SGMMs [4], which lead to improved ASR performance.

The Training Corpus

Recently we collected and manually transcribed a huge training corpus of German broadcast video material, containing 2,705 recordings with a volume of just over 900 hours. The data is segmented into utterances with a mean duration of approximately 5 seconds, yielding 662,170 utterances, and is transcribed manually on word level. The total number of running words is 7,773,971 without taking additional annotation into account. Individual speakers are not annotated, but speaker changes within an utterance are marked and allow for a rough speaker adaptive training scheme. Background noise, speaker noises like breathing and coughing, hesitations, speaker changes, cross-talking speakers, mispronounced words, unintelligible words, and word fragments are annotated by special labels. The recorded data covers a broad selection of news, interviews, talk shows and documentaries, both from television and radio content across several stations.

The data which was used in the baseline systems as described in [1, 6] is a collection of 119,386 utterances with a total duration of 105 hours and 997,996 running words with 62,206 distinct types. Together with the recently transcribed material the corpus has grown to over 1,000 hours of transcribed German broadcast data. All audio is recorded and stored in 16-bit PCM waveform files, with 16 kHz sampling frequency and a single mono channel.

The ASR system

In [7] we used a larger training set and an efficient ASR decoder parameter optimization algorithm i.e. Simultaneous Perturbation Stochastic Approximation (SPSA) [8]. Both approaches led to improved results. The employed training set consisted of 292,133 utterances with a total duration of 322 hours and 3,204,599 running words with 118,891 distinct types. In this work we train models based on recent acoustical modelling algorithms based on DNNs and SGMMs using this training data set. For training of the German LVCSR system we follow the recipes provided in [5]. We further extend the training corpus to a training set consisting of 529,207 utterances with a total duration of 636 hours and 5,940,193 running words with 181,638 distinct types taken from the German broadcast speech corpus described before. We again excluded utterances with background noise, cross-talking speakers, mispronounced words, unintelligible words and word fragments. For decoding we always employ a 3-gram language model with a lexicon consisting of 200k words. For future experiments we will attempt to incorporate even more of the training data. While the corpus is now finished, some of the data was not available at start of the model training.

Evaluation

For evaluation of the LVCSR systems we make use clean speech segments of the DiSCo corpus [1] and use “planned clean speech“(0:55h, 1,364 utterances, 9,184 words) as well as “spontaneous clean speech“(1:55h, 2,861 utterances, 20,740 words). We also evaluate the decoding performance on content from the Rundfunk Berlin-Brandenburg (RBB) provided to the LinkedTV project, again separated into a planned set (1:08h, 787 utterances, 10,984 words) and a spontaneous set (0:44h, 596 utterances, 8,869 words). During annotation of the DiSCo corpus strong emphasis was put on selecting only segments with clean acoustics and without dialectal speech. This was not feasible for the RBB corpus. Hence, especially the spontaneous set partly contains utterances from street interviews with strong dialect.

Using sophisticated acoustical modelling algorithms and larger training data sets we were able to reduce WER on various evaluation sets by a large margin (c.f. table 1). The LVCSR system based on DNNs provides best results when using the large training corpus.

Table 1: WER [%] of different systems on various data sets

System	Size	DiSCo		LinkedTV	
		pl. cl.	sp. cl.	pl.	sp.
GMM [1]	105 h	26.4	33.5	27.0	52.5
GMM [7]	322 h	24.0	31.1	26.4	50.0
DNN	322 h	18.4	22.6	21.2	37.6
SGMM	322 h	18.1	22.5	21.0	36.6
SGMM	636 h	18.1	22.4	20.5	35.9
DNN	636 h	17.4	21.5	19.9	35.3

Topic Segmentation and Labelling

When enriching videos with links to other web resources, proper timing of the topic segments is important to allow for meaningful link expiry. Further, topic labelling is needed when the links are to be filtered and prioritized based on user’s topic preferences. We developed a multi-modal topic segmentation and topic labelling system which will be thoroughly discussed in upcoming publications. However, the ASR output of the described German LVCSR system is employed for subsequent named entity recognition and keyword extraction within this system. Together with a visual shot segmentation approach topic labelling and topic segmentation is performed. Generally the performance of the system improves with higher quality ASR output.

Conclusion and Outlook

The performance of our baseline ASR system [1] was found to be insufficient for subsequent topic segmentation and labelling tasks especially in difficult situations. In this work we presented strategies to improve ASR quality. By increasing training corpus size and by employment of recent acoustic modelling paradigms the WER of the ASR decoder was reduced on various evaluation sets by a large margin. However, we also realize that ASR output from speech data originating from speakers with accents and dialects is still prone to errors. We currently investigate to directly address and incorporate German dialects into our LVCSR system to improve ASR output for dialectal speech.

Acknowledgements

This work has been partly funded by the European Community’s Seventh Framework Programme (FP7-ICT) under grant agreement n° 287911 LinkedTV and n° 269980 AXES.

References

[1] Doris Baum, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim 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.

[2] 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*, 20(1):30–42, 2012.

[3] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006.

[4] Daniel Povey, Lukas Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek, Nagendra K. Goel, Martin KarafiÃ¡t, Ariya Rastrow, Richard C. Rose, Petr Schwarz, and Samuel Thomas. Subspace gaussian mixture models for speech recognition. In *Proc. ICASSP*, pages 4330–4333, 2010.

[5] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[6] D. Schneider, J. Schon, and S. Eickeler. Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System. In *Proc. Association for Computing Machinery’s Special Interest Group Information Retrieval (ACM SIGIR)*, Singapore, 2008.

[7] Jochen Schwenninger, Daniel Stein, and Michael Stadtschnitzer. Automatic parameter tuning and extended training material: Recent advances in the fraunhofer speech recognition system. *Proc. Workshop Audiosignal- und Sprachverarbeitung*, 2013.

[8] James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:3, March 1992.