Exploiting the large-scale German Broadcast Corpus to boost the Fraunhofer IAIS Speech Recognition System

Michael Stadtschnitzer, Jochen Schwenninger, Daniel Stein, Joachim Koehler

Fraunhofer IAIS
Schloss Birlinghoven, 53757 Sankt Augustin, Germany
name.surname@iais.fraunhofer.de

Abstract

In this paper we describe the large-scale German broadcast corpus (GER-TV1000h) containing more than 1,000 hours of transcribed speech data. This corpus is unique in the German language corpora domain and enables significant progress in tuning the acoustic modelling of German large vocabulary continuous speech recognition (LVCSR) systems. The exploitation of this huge broadcast corpus is demonstrated by optimizing and improving the Fraunhofer IAIS speech recognition system. Due to the availability of huge amount of acoustic training data new training strategies are investigated. The performance of the automatic speech recognition (ASR) system is evaluated on several datasets and compared to previously published results. It can be shown that the word error rate (WER) using a larger corpus can be reduced by up to 9.1 % relative. By using both larger corpus and recent training paradigms the WER was reduced by up to 35.8 % relative and below 40 % absolute even for spontaneous dialectal speech in noisy conditions, making the ASR output a useful resource for subsequent tasks like named entity recognition also in difficult acoustic situations.

Keywords: German broadcast speech corpus, Automatic speech recognition, Deep neural networks

1. Introduction and Motivation

The improvement of LVCSR is still in the focus of several research activities in the speech technology area. The WER for spontaneous and natural speech of broadcast recordings for modern ASR systems varies from approximately 15 to 50 % at a rough estimate, depending on the language, the quality of input speech data and speaking styles. Especially when the ASR output is further processed for tasks like named entity recognition it is crucial to use a modern robust ASR system with WERs below 40 % absolute in difficult situations like spontaneous dialectal speech in noisy conditions to make sense at all. Since the LVCSR system which we published in (Baum et al., 2010) does not fulfill this requirement, we describe in this work our activities to improve our German LVCSR system for broadcast data. This system will be continuously improved to achieve comparable result in comparison to leading English based LVCSR. In principle we focused on two ways to improve the acoustic modelling quality of a speech recognition system. First, the availability of larger and extended speech training databases leads to significant and consistent reduction of word error rates. Therefore, we designed a unique German broadcast corpus containing over 1,000 h of transcribed speech data recorded from a variety of German broadcast formats. Second, the application of recent advanced methods for acoustic modelling, for example deep neural networks (DNN) and subspace Gaussian mixture models (SGMM), lead to better ASR performance. In this paper we investigate these two approaches to improve our German broadcast speech recognition system. We show that both approaches, individually and as well as in combination, lead to increased performance of the ASR systems in terms of WER. We also show that the new system provides reasonable results in difficult situations like spontaneous dialectal speech in noisy environments. We evaluate the ASR system on the German Difficult Speech Corpus (DiSCo) (Baum et al., 2010) as well as on data from the LinkedTV[1] project provided by the Rundfunk Berlin Brandenburg (RBB).

2. Related work

LVCSR systems have been successfully employed for various tasks like call center applications, conversational telephone speech transcription, lectures and meetings transcription or speaker-independent automatic broadcast news transcription and indexing. The commercial success of these systems show how far research in LVCSR has come. However, the problem of LVCSR is far from being solved: background noise, accents, dialects, casual and disfluent speech, code switching, and topic changes to name a few, can cause automatic systems to make recognition errors. Nevertheless, technological improvements have been made in all areas of LVCSR i.e. front-end processing, acoustic modelling, language modelling and hypothesis search in the last decade. Mel frequency cepstral coefficients (MFCC) have been widely used in acoustic front-end processing for ASR. Lately, they have been replaced with a more noise-robust representation based on perceptual linear prediction (PLP) coefficients (Hermansky, 1990). Hidden Markov models (Rabiner, 1989) based on state-dependent Gaussian mixture models (GMM-HMM) have been the de facto standard in acoustic modelling for several decades. Many techniques have been proposed to maximize recognition performance for this model. Recently, with advances in DNN pre-training (Hinton et al., 2006), DNN acoustic modelling approaches are reported to outperform discriminatively trained GMM-HMM systems by large amounts (Dahl et al., 2012). Alternative effective training paradigms, like SGMM (Povey et al., 2010), have

[1] http://www.linkedtv.eu

also been proposed. The increased computational power which became available over the last years made it feasible to build ASR systems with several hundreds or even thousands hours of training data.

3. The GER-TV1000h German broadcast corpus

Recently we collected and manually transcribed a huge new training corpus of German broadcast video material, containing 2,705 recordings with a volume of just over 900 h. The new corpus 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. The recorded data covers a broad selection of news, interviews, talk shows and documentaries, both from television and radio content across several stations. This corpus allows us to train much more accurate acoustic models, which will be used for ASR in both the LinkedTV project as well as for the AXES project. In addition to the verbatim transcript, the tags in table 1 were used to further describe the recordings. The tags denote the occurrences of audible background noise, speaker noise, hesitations, speaker changes within an utterance, cross-talking speakers, foreign words, mispronounced words, untranscribable utterances, unintelligible words or word fragments.

The transcribed German broadcast data, which was already used in the baseline system for acoustic model training as described in (Baum et al., 2010; Schneider et al., 2008), is a collection of 119,386 utterances with a total duration of 105 h and 997,996 running words with 62,206 distinct types. This data again covers a broad selection of news, interviews, talk shows and documentaries. Together with the recently transcribed material the corpus has grown to over 1,000 h 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.

4. System configuration

In table 2 an overview of the training sets, which were used for acoustic model training during this work, derived from the GER-TV1000h corpus is given. TS I, the corpus which has already been employed in (Baum et al., 2010; Schneider et al., 2008) consists of 105 h of German broadcast speech and was used to train GMM-HMM based acoustic models. We extend the training corpus size to 322 h (TS II) and train SGMM, DNN and GMM-HMM based models and further extend the training corpus size to 636 h and train SGMM and DNN based acoustic models. Utterances from the recently transcribed material containing one or more annotations with ⟨int⟩, ⟨overlap⟩, ⟨foreign⟩, ⟨mispron⟩, ⟨reject⟩, ⟨**⟩ or ⟨=⟩ were excluded from the training data. Looking at the occurrences of ⟨spk⟩, we observe an average respiratory rate of 9 breaths per minute, and 26 words per breath.

For training of the baseline system (Baum et al., 2010) we used HTK-Toolkit (Young et al., 2006) and Julius (Lee et al., 2001) for decoding. These models are based on generatively trained tri-phone GMM-HMM. For training of the DNN and SGMM based acoustic models we use the training recipes from Kaldi Toolkit (Povey et al., 2011). These models are built upon tri-phone HMM models based on linear discriminant analysis, maximum likelihood linear transformation and speaker adaptive training. The DNN based models use an architecture with four hidden layers each consisting of 1,024 neurons. For decoding we use a 3-gram language model with a pronunciation lexicon consisting of 200,000 words.

The largest training set (TS III) for acoustic model training used in this work consists of 636 h of German broadcast data taken from the GER-TV1000h corpus, which is introduced within this paper. For future works we are still able to further extend the training corpus size using this corpus. The work on the corpus including transcription is already finished, but at time of training start of the acoustic models only parts of the corpus were available, therefore we could not use the whole corpus to create acoustic models for evaluation purposes within this work.

5. Evaluation

An overview of all development and evaluation sets which were used during this work is given in table 3. For development, we use a corpus of German broadcast shows which contains a mix of planned (i.e., read news) and spontaneous (i.e., interviews) speech, for a total of 2,348 utterances (3:29 h, 33,744 words).

For evaluation, we make use of clean speech segments of the DiSCo corpus as described in (Baum et al., 2010), and use “planned clean speech” (0:55 h, 1,364 utterances, 9,184 words) as well as “spontaneous clean speech” (1:55 h, 2,861 utterances, 20,740 words).

Additionally, we test the decoding performance on content from the RBB provided to the LinkedTV project, again separated into a planned set (1:08 h, 787 utterances, 10,984 words) and a spontaneous set (0:44 h, 596 utterances, 8,869 words). While during the annotation of the DiSCo corpus a strong emphasis was put on selecting only segments with clean acoustics (i.e. no background music, high quality) and without dialectal speech, this was not feasible for the newer RBB corpus. Therefore, especially the spontaneous set contains utterances from street interviews, partly with strong Berlin dialect (e.g. “dit Jahr war ja nich berauschend bei Hertha wa” instead of “das Jahr war nicht sehr berauschend bei Hertha”).

In table 4 the results of ASR decoder performance on various German broadcast evaluation sets is given. The system which we published in (Schneider et al., 2008) and further evaluated in (Baum et al., 2010), which is based on generatively trained GMM-HMMs, will serve as baseline configuration in this work. With use of an extended training corpus (TS II, 322 h) the baseline ASR system was improved in terms of WER for all evaluation sets and up to 9.1 % relative. In (Schwenninger et al., 2013; Stein et al., 2013) we further improved the ASR system via decoder parameter optimization based on simultaneous perturbation stochastic approximation (SPSA). This is the only case in this work, were the development corpus was used

label	description
⟨int⟩	If an utterance contains clearly audible background noises, it is tagged with ⟨int⟩. The type and volume of noise was not differentiated in this annotation sequence.
⟨spk⟩	This tag denotes various speaker noises, such as breathing, throat clearing or coughing.
⟨fil⟩	All kinds of hesitations are labelled with this tag.
⟨spk_change⟩	If the speaker changes during the utterance, ⟨spk_change⟩ is inserted at the corresponding position. Using this annotation, speaker turns can be inferred and then used for speaker-adaptive training schemes in later steps.
⟨overlap⟩	If more than one speaker is talking at the same time, the utterance is marked with this tag.
⟨foreign⟩	One or more foreign words, sometimes proper names but most of the time original material with a voice-over.
⟨mispron⟩WORD⟨mispron⟩	Clearly mispronounced words are enclosed in this tag.
⟨reject⟩	If a whole utterance can not be transcribed, it is marked with this tag.
**	If one or more words are unintelligible (e.g. due to background noise), they are transcribed with **.
=	Word fragments are transcribed and end with =, marking them as incomplete.

Table 1: Complete list of labels used for the annotation of GER-TV1000h corpus

Training Set	Duration (h)	# Utterances	# Words	# Words
			total	unique
TS I (Baseline cf. (Baum et al., 2010))	105	119.386	997.996	62.206
TS II	322	292.133	3.204.599	118.891
TS III	636	529.207	5.940.193	181.638

Table 2: Training sets for acoustic modelling derived from GER-TV1000h Corpus

for development purposes i.e. fine-tuning of the decoder parameters. In the other configurations the development set was used for evaluation purposes only. By applying recent modern ASR training paradigms i.e. SGMM and DNN the WER on all experiments was improved by comparatively large amounts for all evaluation sets. The use of SGMM training slightly outperformed DNN training when comparing the models trained on TS II. By employing TS III, the WER of the LVCSR system could be again reduced by approximately 1 % WER absolute for the DNN-based approach. The SGMM-based approach was only slightly improved by employing TS III instead of TS II. Overall, the DNN based approach using the largest training set available i.e. TS III provides the best results. Compared to the baseline system using TS I, the WER could be improved by 7.1 to 17.2 % WER absolute, depending on the evaluation set considered, by using a larger training set and DNN-based models. Without providing exact numbers we noticed that the training of DNN-models is much more time expensive compared to SGMM based models, but during decoding it turns out that DNN decoder is faster compared to SGMM decoder. Results on planned speech are consistently better than results on spontaneous speech as expected. Since the datasets from DiSCo corpus used in this evaluation contain only clean segments and the evaluation sets from LinkedTV contain all types of background noise i.e. there is no categorization of noise types for the LinkedTV corpus, results on DiSCo corpus are better compared to LinkedTV corpus as expected.

6. Conclusion

In this work we presented the large-scale German broadcast corpus (GER-TV1000h) which is composed of over 1000 h of transcribed German broadcast data and which we already partly employed for acoustic model training for German LVCSR. Both the use of larger quantities of speech data and the use of current ASR training paradigms improved the results of the German LVCSR system on various evaluation sets, both individually and together as well. DNN based models using the largest training set (TS III, 636 h) provide the best results for every evaluation set used in this work. In the future we will investigate the usage and exploitation of the full GER-TV1000h corpus to further improve the ASR results for planned and spontaneous speech. This was not feasible during this work because only parts of the corpus were available at begin of the acoustic model training. Compared to the system published (Baum et al., 2010) which serves as a baseline system during this evaluation the results in terms of WER could be improved by 7.1 to 17.2 % WER absolute depending on the choice of the evaluation set. It turns out that even in difficult situation the LVCSR system now produces suitable ASR output for subsequent tasks like named entity recognition or topic segmentation and clustering.

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.

Dataset	Type	Duration (h)	# Utterances	# Words
Development	Development	3:29	2.348	33.744
DiSCo (Baum et al., 2010)	planned clean	0:55	1.364	9.184
DiSCo	spontaneous clean	1:55	2.861	20.780
LinkedTV	planned	1:08	787	10.984
LinkedTV	spontaneous	0:44	596	8.869

Table 3: Development set and evaluation sets

Method	Training Set	WER	WER	WER	WER	WER
		Dev.	DiSCo	DiSCo	LinkedTV	LinkedTV
			planned	spont.	planned	spont.
GMM-HMM (Baum et al., 2010)	TS I	30.2	26.4	33.5	27.0	52.5
GMM-HMM	TS II	29.6	24.0	31.1	26.4	50.0
GMM-HMM, SPSA (Schwenninger et al., 2013)	TS II	27.7	22.6	28.4	24.5	45.6
DNN	TS II	23.9	18.4	22.6	21.2	37.6
SGMM	TS II	23.5	18.1	22.5	21.0	36.6
SGMM	TS III	23.3	18.1	22.4	20.5	35.9
DNN	**TS III**	**22.7**	**17.4**	**21.5**	**19.9**	**35.3**

Table 4: WER [%] results of ASR system configurations on development set and various evaluation sets

7. References

Baum, Doris, Schneider, Daniel, Bardeli, Rolf, Schwenninger, Jochen, Samlowski, Barbara, Winkler, Thomas, and Köhler, Joachim. (2010). 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.

Dahl, George, Yu, Dong, Deng, Li, and Acero, Alex. (2012). Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Processing*, 20(1):30–42.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752.

Hinton, Geoffrey E., Osindero, Simon, and Teh, Yee-Whye. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554.

Lee, A., Kawahara, T., and Shikano, K. (2001). Julius – an Open Source Real-Time Large Vocabulary Recognition Engine. In *Proceedings of Eurospeech*, pages 1691–1694, Aalborg, Denmark.

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

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

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 77(2):257–286.

Schneider, Daniel, Schon, Jochen, and Eickeler, Stefan. (2008). 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.

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

Stein, Daniel, Schwenninger, Jochen, and Stadtschnitzer, Michael. (2013). Improved speed and quality for automatic speech recognition using simultaneous perturbation stochastic approximation. In *Proc. Interspeech*.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.