Open Source German Distant Speech Recognition: Corpus and Acoustic Model

Stephan Radeck-Arneth[1,2], Benjamin Milde[1], Arvid Lange[1,2], Evandro Gouvêa,
Stefan Radomski[1], Max Mühlhäuser[1], and Chris Biemann[1]

[1]Language Technology Group / [2]Telecooperation Group
Computer Science Departement
Technische Universität Darmstadt
{stephan.radeck-arneth,milde,biem}@cs.tu-darmstadt.de
{radomski,max}@tk.informatik.tu-darmstadt.de
{egouvea,arvidjl}@gmail.com

Abstract. We present a new freely available corpus for German distant speech recognition and report speaker-independent word error rate (WER) results for two open source speech recognizers trained on this corpus. The corpus has been recorded in a controlled environment with three different microphones at a distance of one meter. It comprises 180 different speakers with a total of 36 hours of audio recordings. We show recognition results with the open source toolkit Kaldi (20.5% WER) and PocketSphinx (39.6% WER) and make a complete open source solution for German distant speech recognition possible.

Keywords: German speech recognition, open source, speech corpus, distant speech recognition, speaker-independent

1 Introduction

In this paper, we present a new open source corpus for distant microphone recordings of broadcast-like speech with sentence-level transcriptions. We evaluate the corpus with standard word error rate (WER) for different acoustic models, trained with both Kaldi[1] and PocketSphinx[2]. While similar corpora already exist for the German language (see Table 1), we placed a particular focus on open access by using a permissive CC-BY license and ensured a high quality of (1) the audio recordings, by conducting the recordings in a controlled environment with different types of microphones; and (2) the hand verified accompanying transcriptions.

Each utterance in our corpus was simultaneously recorded over different microphones. We recorded audio from a sizable number of speakers, targeting speaker independent acoustic modeling. With a dictionary size of 44.8k words and the best Kaldi model, we are able to achieve a word error rate (WER) of 20.5%.

1.1 Related Work

Table 1 shows an overview of current major German speech corpora. Between 1993 and 2000 the Verbmobil [4] project collected around 180 hours of speech for speech translation. The PhonDat [5] corpus was generated during the Verbmobil project and includes

Table 1. Major available corpora containing spoken utterances for the German language.

Name	**Type**	**Size**	**Recorded**
SmartKom [3]	spontaneous, orthographic & prosodic Transcription	12h	synchronized directional & beamformed
Verbmobil [4]	spontaneous, orthographic transcription	appr. 180h	synchronized close-range & far-field & telephone
PhonDat [5]	spontaneous, orthographic transcription	9.5h	close-range
German Today [6]	read and spontaneous, orthographic transcription	1000h	single microphone, headset
FAU IISAH [7]	spontaneous	3.5h	synchronized close-range to far-field
Voxforge	read, orthographic transcription	appr. 55h	varying microphones, usually headsets
Alcohol Language Corpus [8]	read, spontaneous, command, orthographic transcription & control	300k words	close-range, headset
GER-TV1000h [9]	read, orthographic transcription	appr. 1000h	various microphones, broadcast data
Our speech corpus	read and semi-spontaneous, orthographic transcription	appr. 36h x 3	synchronized multiple microphones far-field & beamformed

recorded dialog speech with a length of 9.5 hours. The Smartkom [3] project combines speech, gesture and mimics for a multimodal application. The data were recorded during Wizard-of-Oz experiments with a length of 4.5 minutes each. Audio was also recorded using multiple far-field microphones, but totals only 12 hours of speech data. A large German speech corpus focusing on dialect identification [6] was recorded by the "Institut für Deutsche Sprache". It contains 1000h of audio recorded in several cities in Germany, thereby ensuring a variety of different dialects. The recorded speakers were split between a group aged between 50-60 years old and a younger group between 16-20 years old. A further speech corpus is FAU IISAH [7] with 3 hours and 27 minutes of spontaneous speech.

The Voxforge corpus [1] was a first open source German speech corpus, with 55 hours of collected speech from various participants, who usually recorded the speech on their own. None of these other German corpora, except the Voxforge corpus, are available under a permissive open source license. However, Voxforge is recorded under uncontrolled conditions, and the audio recording quality is unreliable.

An introduction to distant speech recognition (DSR) is given in [10]. A key challenge is the more pronounced effects, such as noise and reverberation, that the environment has on the recorded speech. Kaldi was recently compared [11] to other open source speech recognizers, outperforming them by a large margin in German and English automatic speech recognition (ASR) tasks. Previously, Morbini et al. [12] also compared Kaldi and PocketSphinx to commercial cloud based ASR providers.

2 Corpus

In this section, we detail the corpus recording procedure and characterize the corpus quantitatively. The goal of our corpus acquisition efforts is to compile a data collection to build speaker-independent distant speech recognition. Our target use case is distant speech recognition in business meetings as a building block for automatic transcription [13] or the creation of semantic maps [14]. We recorded our speech data as

[1] `http://www.voxforge.org`

described in [15]. We employed the KisRecord[2] software, which supports concurrent recording with multiple microphones. Speakers were presented with text on a screen, one sentence at a time, and were asked to read the text aloud. While the setup is somewhat artificial in that reading differs from speaking freely, we avoid the need of transcribing the audio and thus follow a more cost-effective approach.

For the training part of the corpus, sentences were drawn randomly from three text sources: a set of 175 sentences from the German Wikipedia, a set of 567 utterances from the German section of the European Parliament transcriptions [16] and 177 short commands for command-and-control settings. Text sources were chosen because of their free licenses, allowing redistribution of derivatives without restrictions. Test and development sets were recorded at a later date, with new sentences and new speakers. These have 1028 and 1085 unique sentences from Wikipedia, European Parliament transcriptions and crawled German sentences from the Internet, distributed equally per speaker. The crawled German sentences were collected randomly with a focused crawler [17], and were only selected from sentences encountered between quotation marks, which exhibit textual content more typical of direct speech. Unlike in the training set, where multiple speakers read the same sentences, every sentence recorded in the test and development set is unique, for evaluation purposes.

The distance between speaker and microphones was chosen to be one meter, which seems realistic for a business meeting setup where microphones could e.g. be located on a meeting table. There are many more use cases, where a distance of one meter is a sensible choice, e.g. in-car speech recognition systems, smart living rooms or plenary sessions. The entire corpus is available for the following microphones: Microsoft Kinect, Yamaha PSG-01S, and Samson C01U. The most promising and interesting microphone is the Microsoft Kinect, which is in fact a small microphone array and supports speaker direction detection and beamforming, cf. [15]. We recorded the beamformed and the raw signal simultaneously, however due to driver restrictions both are single channel recordings (i.e. it is not possible to apply our own beamforming on the raw signals). We split the overall data set into training, development and test partitions in such a way that speakers or sentences do not overlap across the different sets. The audio data are available in MS Wave format, one file per sentence per microphone. In addition, for each recorded sentence, an XML file is provided that contains the sentence in original and normalized form (cf. Section 3.2), and the speaker metadata, such as gender, age and region of birth.

Table 2 lists the gender and age distribution. Female speakers make up about 30% in all sets and most speakers are aged between 18 and 30 years. For analysis of dialect influence, the sentence metadata also includes the federal state (Bundesland) where the speakers spent the majority of their lives. Despite our corpus being too small for training and testing dialect-specific models, this metadata is collected to support a future training of regional acoustic models. Most speakers are students that grew up and live in Hesse.

The statistics related to the number of sentences per speaker for each of the three text sources in the training corpus is given in Table 3. Most sentences were drawn from Wikipedia. On average, utterances from Europarl are longer than sentences from Wikipedia, and speakers encountered more difficulties in reading Europarl utterances

[2] http://kisrecord.sourceforge.net

Table 2. Speaker gender and age distribution

Gender	**Train**	**Dev**	**Test**
male	105	12	13
female	42	4	4
Age	**Train**	**Dev**	**Test**
41–50	2	0	1
31–40	17	17	2
21–30	108	13	10
18–20	20	1	4

Table 3. Mean and standard deviation of training set sentences read per person per text source and total audio recording times for each microphone

Corpus	**Mean ± StD**
Wikipedia	35 ± 12
Europarl	8 ± 3
Command	18 ± 23
Microphone	**Dur. Train/Dev/Test (h)**
Microsoft Kinect	31 / 2 / 2
Yamaha PSG-01S	33 / 2 / 2
Samson C01U	29 / 2 / 2

because of their length and domain specificity. Recordings were done in sessions of 20 minutes. Most speakers participated in only one session, some speakers (in the training set) took part in two sessions. Table 3 also compares the audio recording lengths of the three main microphones for each dataset. Occasional outages of audio streams occurred during the recording procedure, causing deviations in recording length for each microphone. By ensuring that the training and the dev/test portions have disjoint sets of speakers and textual material, this corpus is perfectly suited for examining approaches to speaker-independent distant speech recognition for German.

3 Experiments

In this section, we show how our corpus can be used to generate a speaker-independent model in PocketSphinx and Kaldi. We then compare both speech recognition toolkits using the same language model and pronunciation dictionary in terms of word error rate (WER) on our heldout data from unseen speakers (development and test utterances). Our Kaldi recipe has been released on Github[3] as a package of scripts, which support fully automatic generation of all acoustic models with automatic downloading and preparation of all needed resources, including phoneme dictionary and language model. This makes the Kaldi results easily reproducible.

3.1 Phoneme dictionary

As our goal is to train a German speech recognizer using only freely available sources, we have not relied on e.g. PHONOLEX[4], which is a very large German pronunciation dictionary with strict licensing. We compiled our own German pronunciation dictionary using all publicly available phoneme lexicons at the Bavarian Archive for speech signals (BAS)[5] and using the MaryTTS [18] LGPL-licensed pronunciation dictionary, which has 26k entries. Some of the publicly available BAS dictionaries, like the one for the Verbmobil [4] corpus, also contain pronunciation variants which were

[3] https://github.com/tudarmstadt-lt/kaldi-tuda-de

[4] https://www.phonetik.uni-muenchen.de/Bas/BasPHONOLEXeng.html

[5] ftp://ftp.bas.uni-muenchen.de/pub/BAS/

included. The final dictionary covers 44.8k unique German words with 70k total entries, with alternate pronunciations for some of the more common words. Stress markers in the phoneme set were grouped with their unstressed equivalents in Kaldi using the `extra_questions.txt` file and were entirely removed for training with CMU Sphinx. Words from the train / development / test sets with missing pronunciation had their pronunciations automatically generated with MaryTTS. This makes the current evaluation recognition task a completely closed vocabulary one and sets the focus of the evaluation on the acoustic model (AM). Still, our pronunciation dictionary is of reasonable size for large-vocabulary speech recognition.

3.2 Language model

We used approximately 8 million German sentences to train our 3-gram language model (LM) using Kneyser-Ney [19] smoothing. We made use of the same resources that were used to select appropriate sentences for recording read speech, but they were carefully filtered, so that sentences from the development and test speech corpus are not included in the LM. We also used 1 million crawled German sentences in quotation marks. Finally, we used MaryTTS [18] to normalize the text to a form that is close to how a reader would speak the sentence, e.g. any numbers and dates have been converted into a canonical text form and any punctuation has been discarded. The final sentence distribution of the text sources is 63.0% Wikipedia, 22.4% Europarl, and 14.6% crawled sentences. The perplexity of our LM is 101.62. We also released both the LM in ARPA format and its training corpus, consisting of eight million filtered and Mary-fied sentences.

3.3 CMU Sphinx acoustic model

Our Sphinx training procedure follows the tutorial scripts provided with the code. We trained a default triphone Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) with Cepstral Mean Normalized (CMN) features. The cepstra uses the standard 13 dimensions (energy + 12) concatenated with the delta and double delta features. The HMM has 2000 senones and 32 Gaussians per state. We further tested the influence of Linear Discriminative Analysis (LDA), Maximum Likelihood Linear Transformation (MLLT) and vocal tract length normalization (VTLN) with a warp window between 0.8 and 1.2 using a step size of 0.02. We used the newest development version of SphinxTrain (revision 12890). SphinxTrain uses Pocketsphinx for the decoding step and sphinx3-align for its forced alignment. After optimization, we ran PocketSphinx with the beamwidth set to 10^{-180} and the language weight to 20.

3.4 Kaldi acoustic models

We follow the typical Kaldi training recipe S5 [1,20] for Subspace Gaussian Mixture Models (SGMM) [21], using a development version of Kaldi (revision 4968). For all GMM models, our features are computed as standard 13-dimensional Cepstral Mean-Variance Normalized (CMVN) Mel-Frequency Cesptral Coefficients (MFCC) features with first and second derivatives. We also apply LDA over a central frame with $+/-$ 4 frames and project the concatenated frames to 40 dimensions, followed by Maximum Likelihood Linear Transform (MLLT) [22]. For speaker adaptation in GMM models we employ feature-space Maximum Likelihood Linear Regression (fMLLR) [23]. We also make use of discriminative training [24] using the minimum phone error rate (MPE) and boosted maximum mutual information (bMMI) [25] criteria. For deep neural network (DNN) - HMM models [26], we also use the standard training recipe with 2048 neurons and 5 hidden layers.

4 Evaluation

Table 4. WER across multiple Kaldi acoustic models.

Kaldi Model	**WER (%)**	
	dev	**test**
GMM	25.7	27.8
GMM+fMLLR	24.6	27.1
GMM+MPE	23.7	26.2
GMM+bMMI(0.1)	23.5	25.8
SGMM+fMLLR	19.6	21.6
SGMM+bMMI(0.1)	19.1	20.9
DNN	**18.2**	**20.5**

Table 5. WER across multiple Sphinx acoustic models.

Sphinx Model	**WER (%)**	
	dev	**test**
GMM	39.6	43.8
GMM+LDA/MLLT	40.5	44.2
GMM+VTLN	**38.3**	**39.6**

Table 4 shows different WER achieved on the development and test sets of our speech corpus, using the Microsoft Kinect speech recordings and different Kaldi models. Table 5 shows our results using different Sphinx models, using the same data resources. For all models the OOV vocabulary of the development corpus was included into the pronunciation dictionary, so the results on this portion reflect the performance of the speech recognizer under a known vocabulary scenario. Adaptation methods like VTLN and fMLLR improve our WER as expected. For Kaldi, using SGMM considerably improves scores compared to purely GMM based models, with a further smaller improvement using a DNN based model. This is probably because the training corpus size is moderate ($\approx$31 hours) and SGMM models usually perform well with smaller amounts of training data [21]. Our best Kaldi model (DNN) clearly outperforms the best Sphinx model (GMM+VTLN) on the test set: 20.5% vs 39.6% WER. Such a relatively large difference in WER performance between the two systems has also been observed in [11]. Domain and training corpus have a large effect on this performance difference [12], but the results presented here do not seem to be unusual for a large vocabulary task in distant speech recognition.

5 Conclusion and future work

In this paper, we present a complete open source solution for German distant speech recognition, using a Microsoft Kinect microphone array and a new distant speech corpus. Distant speech recognition is a challenging task owing to the additional and more pronounced effects of the environment, like noise and reverberation. With a dictionary size of 44.8k words and by using a speaker independent Kaldi acoustic model, we are able to achieve a word error rate (WER) of 20.5%, with a comparatively modest corpus size. We have implicitly exploited the beamforming and noise filtering capabilities of a Microsoft Kinect, which we used, among other microphones for recording our speech data. The Kaldi speech recognition toolkit outperforms the Sphinx toolkit by a large margin and seems to be an overall better choice for the challenges of distant speech recognition.

Ultimately, we would like to enable open source distant speech recognition in German - the presented open source corpus and the acoustic models in this paper is a first step towards this goal. As an extension of our work, we would like to expand our data collection and speech recognition training to study the effects of overlapping and multiple speakers in one utterance, OOV words, additional non-speech audio, and spontaneous speech. This will increase the difficulty of the speech recognition task, but would make it more appealing for research on realistic scenarios. Such scenarios are not uncommon to be in the vicinity of 50% WER or more [27].

Our open source corpus is licensed under a very permissive Creative Commons license and all other resources are also freely available. Unlike other distant speech recognition recipes for German acoustic models, which make extensive use of proprietary speech resources and data with stricter licensing, our resources and acoustic models can be used without restrictions for any purpose: private, academic and even commercial. We also want to encourage the release of other German open source speech data into equally permissive licenses.

Acknowledgements This work was partly supported by the Bundesministerium für Bildung und Forschung (BMBF), Germany within the programme “KMU-innovativ: Mensch-Technik-Interaktion für den demografischen Wandel”.

References

1. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: Proc. IEEE ASRU. (2011) 1–4
2. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I.: PocketSphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices. In: Proc. ICASSP. (2006)
3. Schiel, F., Steininger, S., Türk, U.: The SmartKom Multimodal Corpus at BAS. In: Proc. LREC. (2002)
4. Wahlster, W.: Verbmobil: Translation of Face-To-Face Dialogs. In: Proc. 4th Machine Translation Summit. (1993) 128–135
5. Hess, W.J., Kohler, K.J., Tillmann, H.G.: The Phondat-verbmobil speech corpus. In: Proc. EUROSPEECH. (1995)
6. Brinckmann, C., Kleiner, S., Knöbl, R., Berend, N.: German Today: a really extensive Corpus of Spoken Standard German. In: Proc. LREC. (2008)

7. Spiegl, W., Riedhammer, K., Steidl, S., Nöth, E.: FAU IISAH Corpus – A German Speech Database Consisting of Human-Machine and Human-Human Interaction Acquired by Close-Talking and Far-Distance Microphones. In: Proc. LREC. (2010)
8. Schiel, F., Heinrich, C., Barfüßer, S.: Alcohol language corpus: the first public corpus of alcoholized German speech. Proc. LREC **46**(3) (2012) 503–521
9. Stadtschnitzer, M., Schwenninger, J., Stein, D., Köhler, J.: Exploiting the large-scale German Broadcast Corpus to boost the Fraunhofer IAIS Speech Recognition System. In: Proc. LREC. (2014) 3887–3890
10. Woelfel, M., McDonough, J.: Distant Speech Recognition. Wiley (2009)
11. Gaida, C., Lange, P., Proba, P., Malatawy, A., Suendermann-Oeft, D.: Comparing open-source speech recognition toolkits. `http://suendermann.com/su/pdf/oasis2014.pdf`
12. Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., Traum, D.: Which ASR should I choose for my dialogue system? In: Proc. SIGDIAL. (2013)
13. Akita, Y., Mimura, M., Kawahara, T.: Automatic transcription system for meetings of the japanese. In: Proc. INTERSPEECH. (2009) 84–87
14. Biemann, C., Böhm, C., Heyer, G., Melz, R.: Automatically building concept structures and displaying concept trails for the use in brainstorming sessions and content management systems. In: Proc. I2CS, Guadalajara, Mexico, Springer (2004)
15. Schnelle-Walka, D., Radeck-Arneth, S., Biemann, C., Radomski, S.: An Open Source Corpus and Recording Software for Distant Speech Recognition with the Microsoft Kinect. In: Proc. 11. ITG Fachtagung Sprachkommunikation. (2014)
16. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proc. 10th MT Summit, Phuket, Thailand, AAMT, AAMT (2005) 79–86
17. Remus, S.: Unsupervised relation extraction of in-domain data from focused crawls. In: Proc. Student Research Workshop of EACL, Gothenburg, Sweden (2014) 11–20
18. Schröder, M., Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. IJST **6** (2003) 365–377
19. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Proc. ICASSP. Volume 1. (1995) 181–184
20. Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S., Glass, J.: A complete KALDI recipe for building Arabic speech recognition systems. In: Proc. IEEE SLT, Institute of Electrical and Electronics Engineers Inc. (2015) 525–529
21. Povey, D., Burget, L., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N.K., Karafiat, M., Rastrow, A., Rose, R.C., Schwarz, P., Thomas, S.: Subspace Gaussian Mixture Models for speech recognition. In: Proc. ICASSP. (2010) 4330–4333
22. Gales, M.J.: Semi-Tied Covariance Matrices for Hidden Markov Models. IEEE Trans. Speech and Audio Processing **7** (1999) 272–281
23. Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech & Language **12**(2) (1998) 75–98
24. Gales, M.: Discriminative Models for Speech Recognition. 2007 Information Theory and Applications Workshop (2007)
25. Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., Visweswariah, K.: Boosted MMI for model and feature-space discriminative training. In: Proc. ICASSP. (2008) 4057–4060
26. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech, Language Process. **20**(1) (2012) 30–42
27. Swietojanski, P., Ghoshal, A., Renals, S.: Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In: Proc. IEEE ASRU. (2013) 285–290