Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research

Michael Gref[1], **Joachim Köhler**[1], **Almut Leh**[2]
[1]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, 53757 Sankt Augustin, Germany
[2]Institute for History and Biography, University of Hagen, 58097 Hagen, Germany
{michael.gref, joachim.koehler}@iais.fraunhofer.de, almut.leh@fernuni-hagen.de

Abstract

This paper describes different approaches to improve the transcription and indexing quality of the Fraunhofer IAIS Audio Mining system on Oral History interviews for the Digital Humanities Research. As an essential component of the Audio Mining system, automatic speech recognition faces a lot of difficult challenges when processing Oral History interviews. We aim to overcome these challenges using state-of-the-art automatic speech recognition technology. Different acoustic modeling techniques, like multi-condition training and sophisticated neural networks, are applied to train robust acoustic models. To evaluate the performance of these models on Oral History interviews a German Oral History test-set is presented. This test-set represents the large audio-visual archives "Deutsches Gedächtnis" of the Institute for History and Biography. The combination of the different applied techniques results in a word error rate reduced by 28.3% relative on this test-set compared to the current baseline system while only one eighth of the previous amount of training data is used. In context of these experiments new opportunities are set out for Oral History research offered by Audio Mining. Also the workflow is described used by Audio Mining to process long audio-files to automatically create time-aligned transcriptions.

Keywords: acoustic modeling, robust speech recognition, multi-condition training, speech retrieval, oral history

1. Introduction

The use of automatic speech recognition technology (ASR) to transcribe and index Oral History interviews has started with the MALACH project (Psutka et al., 2002) where the interviews of the *Survivors of the Shoah Visual History Foundation* (VHF) were processed with a state-of-the-art speech recognition in 2002. The main challenge of this activity was the variety and quality of the recordings and the variety of the languages. (Oard, 2012) investigated how speech recognition technology can be used for Oral History research.

In comparison to other speech collections Oral History recordings have a lot challenges. Often the audio recording quality is low. Further the interviewed persons are elderly persons speaking very spontaneous. This leads to poor recognition performance when applying off-the-shelf speech recognition technology.

On the other hand we observe huge progress in speech recognition using different neural network learning frameworks. Recently many researchers, e.g. from Microsoft and IBM, have reported excellent results on conversational recognition tasks, like switchboard. Also with the open-source Kaldi ASR toolkit huge advances are reported using different sophisticated neural network architecture and training methods.

The Fraunhofer IAIS Audio Mining system is designed to automatically create segmented and time-aligned transcriptions from very long, unstructured audiovisual media files. In this work we want to present the opportunities and advantages offered by the Audio Mining system for the Digital Humanities Research by automatically creating transcriptions of Oral History interviews. Moreover we present the challenges for automatic speech recognition systems - such as used in the Audio Mining system - posed by Oral History interviews and approaches to address these challenges.

In this work we apply the Kaldi ASR toolkit to train sophisticated acoustic models for German ASR systems to improve the speech recognition performance on Oral History data. For evaluation we present and apply a novel ASR test-set representing the challenging Oral History data collection of the Institute for History and Biography of the University of Hagen.

The paper is organized as follows. In section 2. the Fraunhofer IAIS Audio Mining system is described. Also the workflow of the audio analysis is presented that automatically segments and transcribes audiovisual media data. The research on Oral History at the Institute for History and Biography is described in section 3. where the main focus is put on the Oral History database. Furthermore we describe the advances and new opportunities offered by the Audio Mining system for Digital Humanities Research using Oral History. In section 4. we describe the challenges that have to be met in order to achieve reasonable results for the automatic transcription of Oral History interviews. We also present the approaches taken to face these challenges. Training and evaluation of sophisticated acoustic models using these approaches for robust speech recognition are presented in section 5.

2. The Fraunhofer IAIS Audio Mining System

2.1. Overview

The Fraunhofer IAIS Audio Mining system is designed to analyze the audio-signal of audiovisual media files automatically. The aim is to create a time-aligned transcription of the spoken words, as it is normally made by professional human transcribers. To achieve an optimal result this does not only include automatic speech recognition but an entire workflow including segmentation of the audio signal, speech detection, speaker analysis and keyword extrac-

Figure 1: Graphical Web User Interface of the Fraunhofer IAIS Audio Mining system

tion using several state-of-the-art pattern recognition algorithms.

Currently, the Fraunhofer IAIS Audio Mining system enables journalists, archivists and hosts of audiovisual broadcast data to face the challenges caused by the continuously increasing amounts of large audiovisual data. This is achieved by making the files both text-searchable and structured. Thus the amount of time a user needs to work with AV-data is noticeably reduced.

For example, the system enables end-users to quickly navigate within interviews using a Graphical User Interface (GUI) that exploits the analysis-results provided by the Audio Mining system. One example for such a GUI using the Fraunhofer IAIS Audio Mining system is shown in figure 1. An embedded media-player allows users to directly play segments of specific speakers. Different speakers are represented by a unique color for each speaker in the time-bar below the video. Non-speech is represented by grey. Furthermore, a search-engine enables the user to find all media files in which a keyword or phrase was spoken by searching the transcripts of spoken words provided by the automatic speech recognition. The GUI highlights all occurrences of the searched words in the time-bar of the currently played media-file.

2.2. Workflow of the Audio Analysis Subsystem

In the following we describe the aforementioned workflow of the *Audio Analysis Subsystem* as component of the Audio Mining system in more detail. This description is based on recent work at the Fraunhofer IAIS Institute (Schmidt et al., 2016). The schematic structure of the audio analysis workflow for one media-file is illustrated in figure 2.

2.2.1. Audio Segmentation

The audio signal is first cut into segments by an audio segmentation algorithm. For this the *Bayesian Information Criterion* (BIC) is applied on full covariance Gaussian models of the mel-frequency cepstral coefficents (MFCC) (Tritschler and Gopinath, 1999).

2.2.2. Concept Detection

After segmentation, each segment is classified using a speech/non-speech detection. Segments that are assumed to contain speech are additionally classified using a gender detection and then passed to the following processing steps. The two detection algorithms are *Gaussian Mixture Model-Universal Background Model* (GMM-UBM) approaches trained for the respective classification task.

2.2.3. Speaker Clustering

The aim of speaker clustering is to classify all speech-segments in which the same speaker is talking. Recently we adapted iVectors (Dehak et al., 2011) for this task and achieved increased performance compared to our previous BIC-based algorithm.

2.2.4. Automatic Speech Recognition

The automatic speech recognition used within the current version of Audio Mining system is trained using the widely adopted Kaldi ASR toolkit (Povey et al., 2011). The currently used acoustic model is a hybrid *Hidden Markov Model-Deep Neural Network* (HMM-DNN) approach with a fully-connected DNN trained on the 1005h large-scale German broadcast corpus *GerTV1000h* (Stadtschnitzer et al., 2014) of the Fraunhofer IAIS Institute.
The language model is trained on a text-corpus of German broadcast data. For the experiments described in this work a

Figure 2: Audio analysis workflow

lexicon with about $500,000$ words is used. However, we recently trained language models for specific tasks with more than one million words in the lexicon.

The ASR system was evaluated on the DiSCo corpus (Baum et al., 2010). DiSCo is a German evaluation corpus for challenging problems in the broadcast domain and is split in four evaluation sets: planned and spontaneous speech each in clean and mixed noise conditions. The system achieves 15.3% word error rate (WER) on clean planned speech and 19.4% WER on clean spontaneous speech. However, as the following sections show, Oral History interviews are far more challenging for the ASR-system. To achieve satisfactory results advanced modeling approaches have to be applied.

2.2.5. Keyword Extraction

In the last step of analysis keywords are extracted from the ASR-generated transcript using a *tf-idf* (term frequency-inverse document frequency) approach. This allows users to search and filter media files that contain a specific topic.

3. Description of Hagen Database

3.1. Oral History Interviews as Sources for the Humanities

Research based on interviews with witnesses to historical events and the interest in biographical processes and subjective personal information have a long tradition in the social sciences and humanities. Since the early 1980s biographical research emerged in almost all areas of the humanities: sociology and pedagogy, ethnography and ethnology, historical and literary studies, as well as in psychoanalysis and psychology. In the historical sciences, research conducting and analyzing interviews with contemporary witnesses has become known as *Oral History*. Particularly in Germany this research was focused above all on the period of National Socialism and the Second World War. But in the meantime, it has also come to include many other topics and historical periods. The past forty years have seen a multitude of witnesses to a wide range of historical events interviewed by researchers. Today it is hard to imagine the presentation of historical information in exhibitions, documentations and films without using witness accounts to the relevant events.

This method in which most of the interviews in question were conducted is characterized by the fact that rather than structuring the interview around questions, the interviewer encourages the interviewee to freely narrate his or her life story. In terms of biographical research, the outcome is qualified as a narrative life-story interview lasting very often at least three or four hours.

Such an interview is representing a highly individual testimony in which the interviewee has presented large parts of his life story and his world view in a way that is often unguarded and sometimes contradictory. Due to the open character of the narration and the life-story dimension such an interview is worth for more than one interpretation and a valuable source for later re-use the more as many witnesses died meanwhile leaving only their recorded account. For the same reason analyzing as well as archiving Oral History interviews is useful and challenging.

Today archives, museums, historical sites and documentation centers preserve and provide Oral History interviews for historical research, social sciences and other humanities.

3.2. The Oral History Archives "Deutsches Gedächtnis"

The archives "Deutsches Gedächtnis" ("German Memory") provide about 2,500 Oral History interviews conducted from 1975 to this day in more than 100 projects using various recording technologies and interview settings. The average length of the interviews is 3.5 hours. The interview is not structured by questions but is open for the course of memories coming into the interviewee's mind when telling his entire life story from birth and childhood into the present. The interviewees were born between 1895 and 1980.

The original analog recordings of the 2,000 audio and 500 video-interviews are digitized. Although retrieval as well as analysis is based on the transcription to date only half of the interviews are transcribed and saved as text files. Only ten percent of the transcripts are time aligned so that the transcript works as subtitle to the audio or video recording. All interviews are equipped with archival, technical and biographical meta-data.

3.3. The ASR-Test-Set

The ASR-test-set is a subset of the "Deutsches Gedächtnis" archives representing the wide range of interviews with respect to recording technology, interview methodology, dialects and pronunciation. The recording quality and the pronunciation had to be understandable for humans as a precondition to be used as data for the test-set. The selection should include early interviews as well as recent conducted ones and represent the interview method of various academic disciplines. With respect to gender and age the selection should display the entire collection. Within these criteria the test-set was randomly selected. The test-set contains 102 audio files from 35 different speakers with an overall length of about 3.5 hours, $27,053$ spoken words and $4,592$ unique words. The recordings used for the test-set took place between 1980 and 2012.

3.4. Advantages Offered by Audio Mining for Interview-Based Research

Audio Mining offers advantages for archiving and retrieval as well as for analysis and interpretation of Oral History in-

terviews. For both, archiving and analyzing, time-aligned transcription and indexing with keywords is essential. Currently transcribing, labeling and annotating speech recordings is performed completely manually. Due to the huge effort in terms of time and human resources required to do this the efficiency to exploit Oral History interviews for digital humanities research is severely limited.

By transcribing audio-visual recordings semi-automatically and providing additional speech related analysis features - indexing and structuring the content - the Audio Mining tools allows to process huge amounts of Oral History data and enhance retrieval and research based on these interviews.

Regarding archiving and retrieval the Audio Mining tools allow full-text search with direct access to the audio/video recording so that search results can be checked immediately.

With respect to the research process new approaches in quantity and quality are made possible. Covering more data comparative studies and quantitative analysis become reasonable. Furthermore the speech technologies allow to analyze verbal and non-verbal aspects of communication more deeply thus opening new dimensions for qualitative research.

From the early days of Oral History to this day oral historians insist on the oral nature of their sources. Following their demand the audio tape or file is the primary source and should be the main subject of research. In consequence the transcript is only a necessary additive for the analysis. In fact and in practice, the transcript is very often the only source for interpretation and analysis, replacing the audio completely. The subtitled audio/video recording overcomes this discrepancy thus fulfilling the essential demand to analyze not only the transcript but the complete oral source. Thus Audio Mining allows to take full advantage of the untapped potential of Oral History leading back to its original roots.

The special challenge of Oral History interviews is determined by spontaneous colloquial speech and the poor technical quality of the audio material recorded decades ago with insufficient equipment.

4. Challenges and Advanced Modeling Approaches

4.1. General State-of-the-Art Acoustic Modeling Approaches

At the Fraunhofer IAIS we started applying deep-learning for acoustic modeling with the well-known (classical) hybrid HMM-DNN approach some years ago using a fully-connected deep neural network. One of these models is still used in the current production system of Audio Mining. In one of our recent previous works (Schmidt et al., 2016) we trained *recurrent neural networks* (RNNs) with *connectionist temporal classification* (CTC) (Graves et al., 2006) as the objective function and the *Eesen-ASR-Toolkit* (Miao et al., 2015) for acoustic modeling achieving increased speech recognition performance. A comparison of word error rates achieved by hybrid HMM-DNN-Models and CTC-RNN-Models we trained in previous works are given in table 1.

However, many recent works (e.g. (Saon et al., 2017) as one of many) show a significant increase in performance on acoustic modeling using the (classical) hybrid HMM-DNN approach with sophisticated neural network architectures. Thus we decided to intensify our research on acoustic modeling using hybrid HMM-DNN models for the German language. For this we train acoustic models with different state-of-the-art network architectures using the Kaldi ASR toolkit. We aim to increase the general performance of the automatic speech recognition system as well on broadcast data as on Oral History interviews and future applications.

4.2. Audio Signal Quality

The audio signal quality is one of the main challenges we are facing when adapting the ASR system to Oral History interviews. The broadcast audio signals of the GerTV1000h corpus used for training the acoustic model were recorded and post-processed using highly professional equipment. The recordings usually have no background noise, barely perceptible reverberation and the levels are well adjusted. However, Oral History interviews are usually recorded in the living rooms of the contemporary witnesses using commonly available recording equipment. This equipment changed during the years of recordings and result in a wide range of different sound qualities: from nearly clean recordings to recordings with noises, reverberation and even clipping.

There are two main approaches how to face these challenges: *Speech Enhancement* and *Multi-Condition Training*. Speech enhancement aims to modify the signal itself and increase the audio quality by reducing noises and compensating corruptions. *Multi-Condition Training* on the other hand aims to make the acoustic model robust against corruptions by showing the model a wide range of features of corrupted audio signals during training. This approach aims to let the model generalize to unseen noise-types and distortions by relying only on robust acoustic features. In this work we utilize the latter approach and try to train such a robust acoustic model.

Therefore we virtual corrupt the GerTV1000h corpus with *data augmentation* as described in subsection 5.4. Defining discrete-time-signals as sequences of sample values the augmentation can be described as

$$(x_n)_{n\in\mathbb{N}} := (s_n)_{n\in\mathbb{N}} * (h_n)_{n\in\mathbb{N}} \tag{1}$$

if only reverberation and no background noise affects the speech signal or as

$$(\tilde{x}_n)_{n\in\mathbb{N}} := (s_n)_{n\in\mathbb{N}} * (h_n)_{n\in\mathbb{N}} + (w_n)_{n\in\mathbb{N}} * (\tilde{h}_n)_{n\in\mathbb{N}} \tag{2}$$

if also noise inside the room affects the speech signal.

Here $*$ is the convolution operation for sequences, $(s_n)_{n\in\mathbb{N}}$ the sequence of the clean speech signal, $(h_n)_{n\in\mathbb{N}}$, $(\tilde{h}_n)_{n\in\mathbb{N}}$ are room-impulse-responses modeling the reverberation of one room at different positions and $(w_n)_{n\in\mathbb{N}}$ the sequence of the noise-signal.

Such approaches are quite common but still state-of-the-art methods to create multi-condition training data when only clean data is available. For example Google recently used a similar approach to train Google Home on far-field, multichannel speech recognition (Li et al., 2017).

4.3. Language

Other big challenges posed by Oral History interviews are colloquial language used in spontaneous speech, hesitations, age- and health-related changes in the way of speaking and domain specific words used in the interviews that usually do not occur in everyday speech (e.g. *Kriegerwitwensöhne*, German for *sons of a war widow*). These challenges must be addressed by adapting the language model and will be part of our future work. In this work we focus on acoustic modeling only and use our default language model described in subsection 2.2.4..

5. Experiments and Evaluation

5.1. Experimental Setup

In the following we use the Oral-History ASR-Test-Set described in subsection 3.3. along with the aforementioned DiSCo test-sets to measure the gain proposed by the approaches described in the previous section.
In the first set of experiments we train hybrid HMM-DNN acoustic models using different types of sophisticated neural networks. These models are build and trained using the *nnet3*-implementations provided within the Kaldi-framework. We run experiments on the following architectures of neural networks:

- **LSTM** (*Long short-term memory* neural networks):
  Long short-term memory is an architecture for recurrent neural networks proposed by (Hochreiter and Schmidhuber, 1997). The Kaldi-nnet3-implementation uses LSTM-Layers with forget-gates (Gers et al., 2000), peephole connections (Gers and Schmidhuber, 2000) and projection layers (Sak et al., 2014). The LSTM-Network configuration we used consists of three LSTM-Layers.

- **BLSTM** (Bidirectional LSTMs):
  Bidirectional LSTMs were first proposed by (Graves and Schmidhuber, 2005). However, the underlying concept of bidirectional recurrent neural networks was proposed by (Schuster and Paliwal, 1997). The neural networks has three BLSTMs-Layers in the configuration we used.

- Chain-Models (Povey et al., 2016):
  The Chain-Models we trained used both LSTM and TDNN-Layers (Waibel et al., 1989), (Peddinti et al., 2015) in one network. We show the results of two Chain-Model configurations:

  - **Chain A**: The network consists of 10 layers (7 TDNN-Layer and 3 LSTM-Layer). See (Cheng et al., 2017)) for a detailed description of the setup.
  - **Chain B**: The network architecture is equal to **Chain A** except for the usage of the (default) LSTM-Layer-implementation *LSTMp* instead of *FastLSTMp*, the application of per-frame-dropout, as described in (Cheng et al., 2017), and minor training parameters changes.

The LSTM-Layers in all networks have a cell-dimension of 1024 and a projection-dimension of 256 for unidirectional topologies. In bidirectional networks a projection-dimension of 128 is used.
The models are trained on 40-dimensional MFCC features with 5 consecutive frames at each time-step append with a 100-dimensional iVector (Dehak et al., 2011). This gives a 300-dimensional input at each time-step.
As in our previous works we train the acoustic models on the GerTV1000h corpus. However, due to the computational time needed for training sophisticated neural networks we decided to use a 128 h subset of the GerTV1000h corpus for the experiments in this work. For each experiment the development set (Dev Set) is used to adjust the Language-Model-weight to a fixed value across all test-sets within the respective experiment.

5.2. Previous Results

In our previous work we focused on speech recorded in the broadcast-domain. A detailed description of this work is given in (Schmidt et al., 2016). As described before, recordings in the broadcast-domain can be considered very clean. Therefore, for evaluation we mostly used the clean subset of the DiSCo corpus in our previous work. The word error rates of different acoustic models trained in our previous work are summarized in table 1. All these models were trained on the entire GerTV1000h corpus. For evaluation the same language model for evaluation was used - except for the last row in the table, where we applied a lower pruning factor (*Prun. Fact.*) on the language model.

Model	**Prun. Fact.**	**Dev Set**	**DiSCo clean**	
			plan.	spont.
HMM-DNN*	1e-7	21.3	15.3	19.4
HMM-pDNN	1e-7	18.8	13.3	16.5
CTC-RNN	1e-7	18.1	12.8	15.4
CTC-RNN	1e-8	**17.2**	**11.9**	**14.5**

Table 1: Overview of the word error rates of our previous acoustic modeling approaches trained on the entire GerTV1000h Corpus

The HMM-DNN-model marked with * is the acoustic model used in the current production system we aim to improve. This baseline-system gives a 55.0% word error rate on the Oral History test-set.

5.3. Training State-of-the-Art Acoustic Models On a 128 h Training-Data Subset

First we analyze the influence of using the previously described sophisticated neural network architectures for the hybrid HMM-DNN acoustic modeling approaches. We trained all models on a 128 h subset of the GerTV1000h corpus to save computational time. All experiments were performed using the same 500,000 words language model that was already used in our previous works. We used 1e-8 as the pruning factor since we achieved best results with this configuration in our previous work. The word error rates for all test-sets are summarized in table 2.

Model	Dev Set	DiSCo clean		DiSCo mixed		Oral Hist.
		plan.	spont.	plan.	spont.	
LSTM	16.6	12.1	13.9	16.7	26.0	**45.3**
BLSTM	17.2	12.8	14.6	17.6	27.6	48.2
Chain A	17.0	11.6	13.8	17.4	28.9	53.4
Chain B	**15.5**	**10.6**	**12.4**	**15.1**	**25.0**	50.2

Table 2: Overview of the word error rates using sophisticated acoustic modeling approaches trained on a 128 h subset of the GerTV1000h corpus

The best results on the DiSCo test-sets are achieved using the Chain B model. Both Chain models outperform our previously best model (Table 1, CTC-RNN with 1e-8 PF) on the DiSCo clean planned and spontaneous test-sets, even though the Chain models were only trained on a 128 h subset of the GerTV1000h corpus.
However, the Chain models perform rather bad on the Oral History test-set compared to the simpler HMM-LSTM-Models. One possible reason could be some kind of overfitting of the Chain models to the clean training data. The unidirectional LSTM-Model achieves the best word error rate on the Oral History test-set (45.3%) reducing the word error rate by 9.7% absolute and 17.6% relative compared to the baseline system.
Surprisingly, the bidirectional LSTM model performs worse than the unidirectional LSTM on all test-sets. This could be a lack of generalization during training due to the reduced size of training data using the 128 h subset. However since both the LSTM and Chain B model perform better than the BLSTM, we neglect the BLSTM in the following experiments.

5.4. Training Robust Acoustic Models On 128h Multi-Condition Data

With the following set of experiments we analyze the influence of multi-condition training on robustness of the acoustic models. These experiments are all carried out training the LSTM acoustic model setup that achieved best results on the Oral-History test-set in the previous experiments. We assume that the results of these multi-condition training generalize to other network architectures and training on the entire GerTV1000h Corpus as well. This will be analyzed in future our work.
We apply the data augmentation techniques described in section 4.2. to artificially distort the utterances of the GerTV1000h corpus creating three noisy variants of the corpus:

- **Reverb**: All signals are convolved according to equation (1) with randomly selected room impulse responses of small or medium-sized rooms. No noise is applied here.
- **R+AWGN**: Similar to **Reverb** but added white Gaussian noise to the signals.
- **R+RealNoise**: Similar to **R+AWGN** but instead of AWGN non-stationary noises are added that have been randomly selected from real-life recordings, e.g. street noises, noises in a bus, police sirens, hairdryers.

The noises in **AWGN** and **RealNoise** were also convolved with a room-impulse-response before superposing them with the reverberant speech signal. According to equation (2) we used a different room-impulse-response of the same room, than the one applied to the speech-signal, if one was available. Otherwise we used the same room impulse response. A random signal-to-noise ratio between 10 and 20 was applied for the superposition.
We used 266 room-impulse-responses of small and medium-sized rooms collected from different sources. The real-life noises contain 14.5 h of recordings. To avoid overfitting to the noises we superposed up to three different randomly selected noises for one audio file.
In our multi-condition experiments we want to analyze the influence of training acoustic models using different mixtures of the **Clean**, **Reverb**, **R+AWGN** and **R+RealNoise** data-sets. Thus the data-sets we actually use for training are created by randomly selecting each file from one of the four different data-sets using one of the distributions set out in table 3. This way we created four different multi-condition training-sets (v1 to v4) beside the clean one. For the 128 h subset we then selected the same utterances that were used in the previous (clean) experiments.

Conf.	Clean	Reverb	R+AWGN	R+RealNoise
Clean	100 %	0 %	0 %	0 %
v1	50 %	50 %	0 %	0 %
v2	40 %	40 %	20 %	0 %
v3	35 %	35 %	15 %	15 %
v4	40 %	40 %	0 %	20 %

Table 3: Overview of the different multi-condition-training configurations used for the GerTV1000h Corpus

Conf.	Dev Set	DiSCo clean		DiSCo mixed		Oral Hist.
		plan.	spont.	plan.	spont.	
Clean	**16.6**	**12.1**	**13.9**	**16.7**	**26.0**	45.3
v1	17.1	12.4	14.4	16.9	27.0	40.3
v2	17.3	12.7	14.5	16.8	26.7	39.7
v3	17.4	12.8	15.1	17.3	26.8	40.1
v4	17.3	12.4	14.6	16.7	26.5	**39.4**

Table 4: Word error rates of an LSTM-Model trained on 128 h training data using different multi-condition configurations

The word error rates of all test-sets and all models trained on the different multi-condition training-data-sets are summarized in table 4. It is very promising that all LSTMs generalized well being trained on the different presented configurations. On the one hand the difference of word error rates on the Oral History test-set between v1, v2, v3 and v4 is below 1% absolute. On the other hand the results on all four DiSCo test-sets remain well compared to the clean-

trained model. For v4 the word error rate on DiSCo increases by 0.7% absolute in the worst case while achieving the best result on the Oral History test-set (39.4%). With this configuration we were able to reduce the word error rate on the Oral History test-set furthermore by 5.9% absolute compared to the clean trained LSTM. Overall we improved the word error rate by 15.6% absolute and 28.3% relative compared to the baseline-system.

6. Conclusion and Outlook

In this work we presented the opportunities and advantages offered by automatic speech recognition systems, such as the Fraunhofer IAIS Audio Mining system, for Digital Humanities Research by automatically creating segmented and time-aligned transcriptions of Oral History interviews. We also present the challenges caused by Oral History interviews for ASR-systems. State-of-the-art acoustic modeling approaches such as sophisticated neural network architectures and multi-condition training were applied to cope with these challenges. Evaluated the systems on our proposed Oral History ASR test-set, we were able to improve the word error rate by 15.6% absolute and 28.3% relative compared to the baseline-system while only one eighth (128 h) of the previous amount of the training data was used.

We plan to utilize the Fraunhofer IAIS Audio Mining system to automatically process all not-transcribed interviews of the archives "Deutsches Gedächtnis" of the Institute for History and Biography. Future user studies need to show if the automatic transcription can furthermore be used as a basis for creating a perfect (time-aligned) transcript by human error-correction in the productive use.

However, to achieve satisfactory results the speech recognition performance has to be improved further. Therefore our future work aims to further decrease the word error on Oral History interviews. We want to further improve robust acoustic modeling, e.g. by training models on the entire amount of the GerTV1000h training data and run experiments to further increase the amount of training-data by a factor of 3 or 4 using the proposed data-augmentation methods. Moreover we plan to train domain-specific language models for Oral History interviews.

We also plan to time-align the manually transcribed but not time-aligned interviews of the "Deutsches Gedächtnis" archives using *forced alignment*-approaches for very long audio-files. This would provide new opportunities for Oral History research and could also be utilized for acoustic model training.

Acknowledgement

This research has been funded by the Federal Ministry of Education and Research of Germany (BMBF) in the project *KA*^{3} *- Kölner Zentrum für Analyse und Archivierung von AV-Daten* (Cologne center for the analysis and archiving of audiovisual data) (project number: 01UG1511B).

7. Bibliographical References

Cheng, G., Peddinti, V., Povey, D., Manohar, V., Khudanpur, S., and Yan, Y. (2017). An exploration of dropout with lstms. In *Proc. Interspeech 2017*, pages 1586–1590.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May.

Gers, F. A. and Schmidhuber, J. (2000). Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pages 189–194 vol.3.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471, Oct.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4, July.

Graves, A., Fernández, S., and Gomez, F. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *In Proceedings of the International Conference on Machine Learning, ICML 2006*, pages 369–376.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780, November.

Li, B., Sainath, T. N., Narayanan, A., Caroselli, J., Bacchiani, M., Misra, A., Shafran, I., Sak, H., Pundak, G., Chin, K., Sim, K. C., Weiss, R. J., Wilson, K. W., Variani, E., Kim, C., Siohan, O., Weintraub, M., McDermott, E., Rose, R., and Shannon, M. (2017). Acoustic modeling for google home. In *Proc. Interspeech 2017*, pages 399–403.

Miao, Y., Gowayyed, M., and Metze, F. (2015). Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174, Dec.

Oard, D. (2012). Can automatic speech recognition replace manual transcription? In Doug Boyd, et al., editors, *Oral History in the Digital Age*, Washington, D.C. Institute of Museum and Library Services.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 3214–3218.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (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.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech 2016*, pages

2751–2755.

Psutka, J., Ircing, P., Psutka, J. V., Radová, V., Byrne, W. J., Hajič, J., Gustman, S., and Ramabhadran, B. (2002). Automatic transcription of czech language oral history in the malach project: Resources and initial experiments. In Petr Sojka, et al., editors, *Text, Speech and Dialogue: 5th International Conference, TSD 2002 Brno, Czech Republic, September 9–12, 2002 Proceedings*, pages 253–260, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sak, H., Senior, A. W., and Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, abs/1402.1128.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., and Hall, P. (2017). English conversational telephone speech recognition by humans and machines. In *Proc. Interspeech 2017*, pages 132–136.

Schmidt, C. A., Stadtschnitzer, M., and Köhler, J. (2016). The fraunhofer iais audio mining system: Current state and future directions. *Speech Communication; 12. ITG Symposium, Paderborn, Germany, 2016*, 12:115–119.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, Nov.

Tritschler, A. and Gopinath, R. (1999). Improved speaker segmentation and segments clustering using the bayesian information criterion. In *EUROSPEECH'99*, Budapest, Hungary, september. ISCA.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, Mar.

8. Language Resource References

Baum, Doris and Schneider, Daniel and Bardeli, Rolf and Schwenninger, Jochen and Samlowski, Barbara and , Winkler Thomas and Köhler, Joachim. (2010). *DiSCo - A German Evaluation Corpus for Challenging Problems in the Broadcast Domain*. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Stadtschnitzer, Michael and Schwenninger, Jochen and Stein, Daniel and Köhler, Joachim. (2014). *Exploiting the Large-Scale German Broadcast Corpus to Boost the Fraunhofer IAIS Speech Recognition System*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).